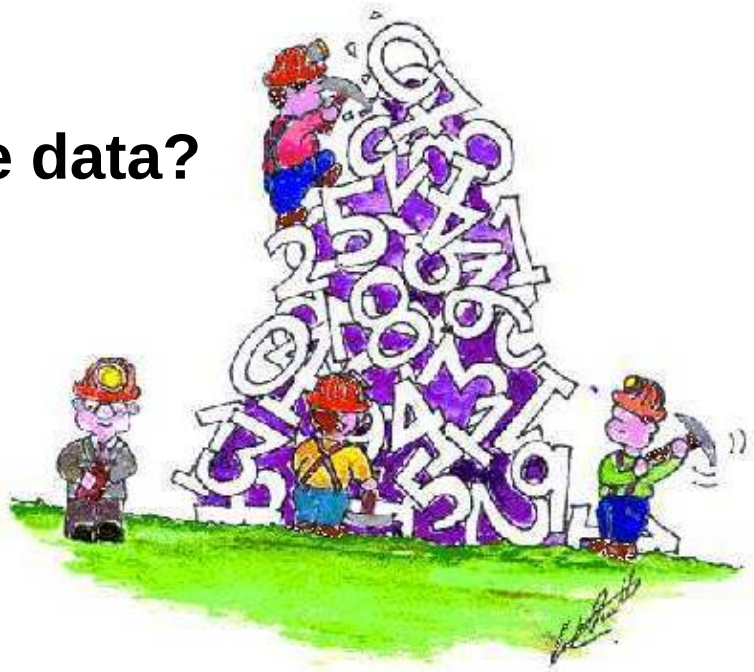
Applied Machine Learning Course

Pitfalls and Caveats in Machine Learning

What matters most?

the data?



the algorithm?

Both matter!

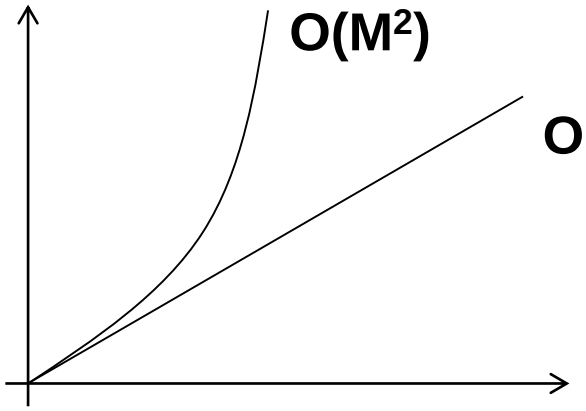
Gathering data & the curse of dimensionality

The more data the merrier!

Not always true

For most machine learning algorithms, computation costs grow proportionally to the dimension and number of datapoints.

Computational Costs



Linear or exponential growth?

Properly choosing the dataset

Example: Classify emails into spam and not-spam

Information you can use:

IP-address **@epfl.ch** **.edu** **.org** **.com**

Some may be valid emails

Reply-to address **noreply@example.com**

sender-address@*example*.com

Some may look perfectly valid
Careful understanding of the content can tell them apart

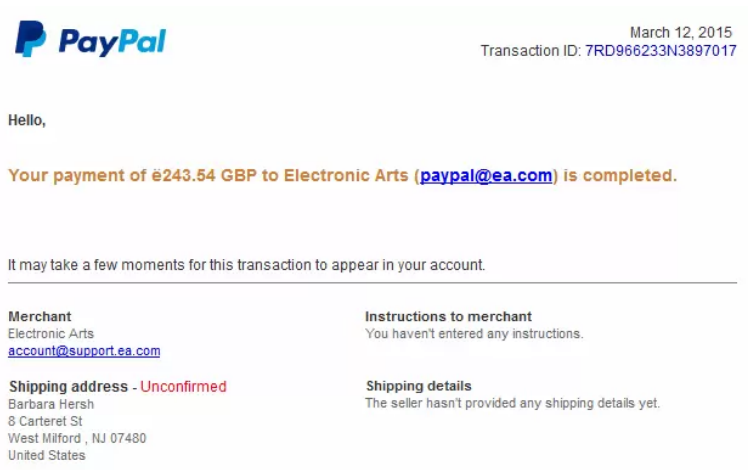
Subject line **Your payment is ready**

Email quota reached

Unsubscribe

WINNING AMOUNT

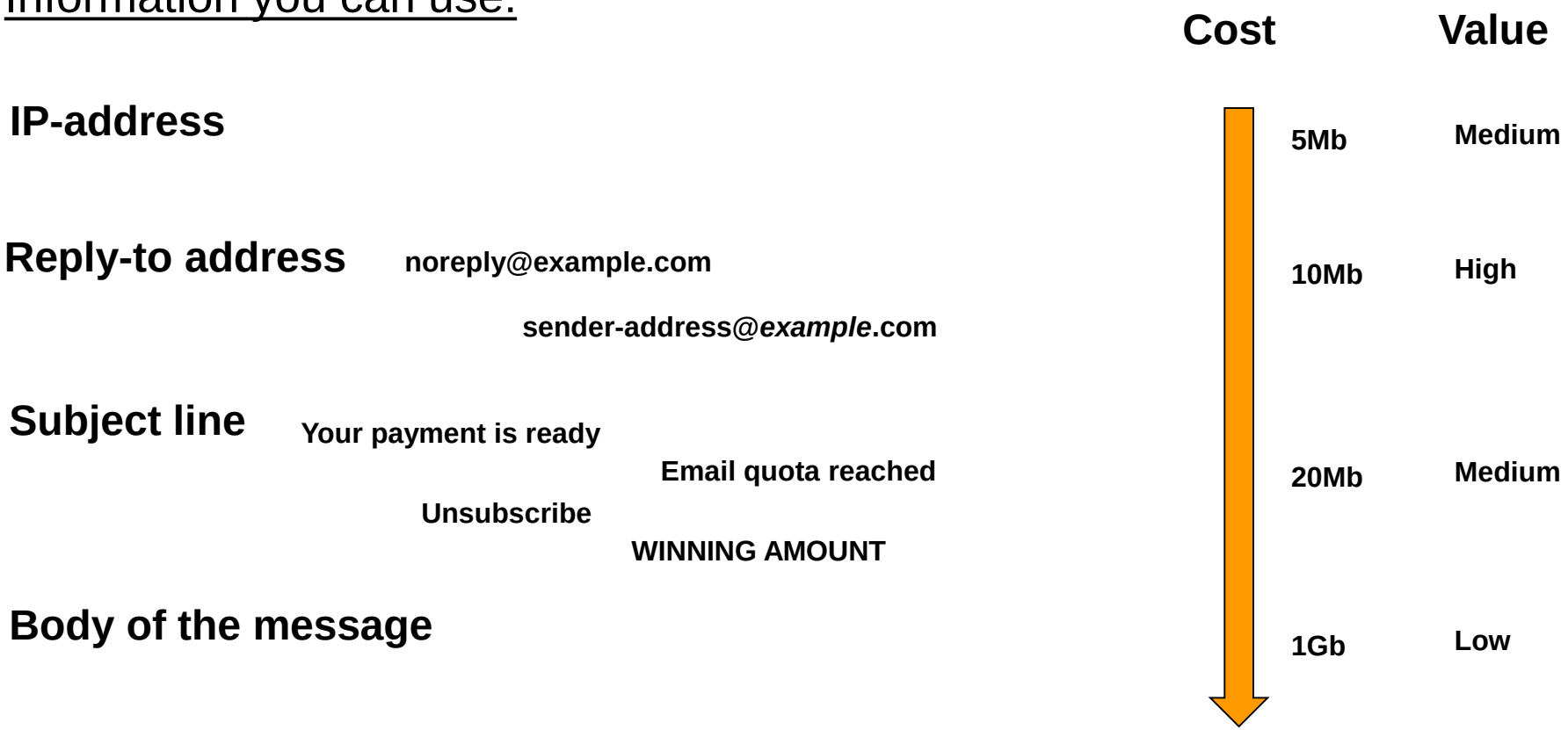
Body of the message



Properly choosing the dataset

Example: Classify emails into spam and not-spam

Information you can use:



Properly choosing the dataset

The more data the merrier!

Not always true

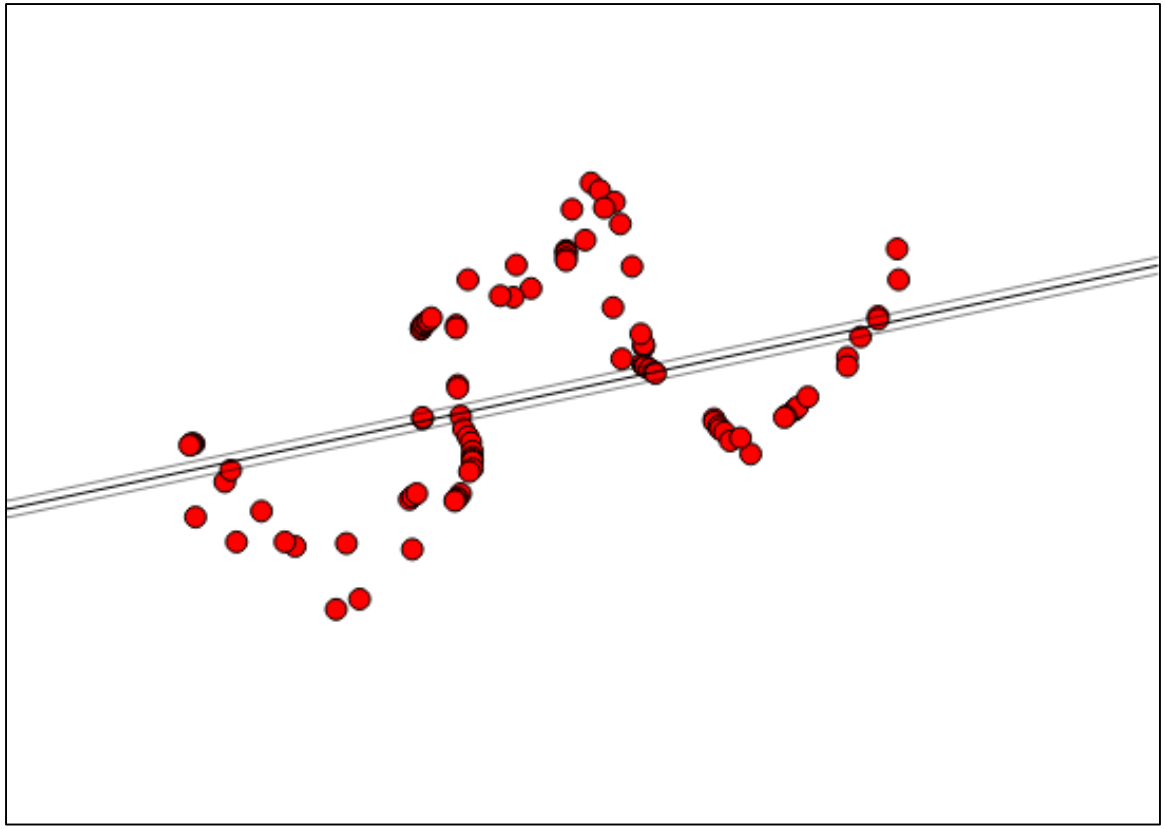
Machine learning is heavily based on statistics.

More data may introduce noise and make it harder for the algorithm to tell apart what is relevant.

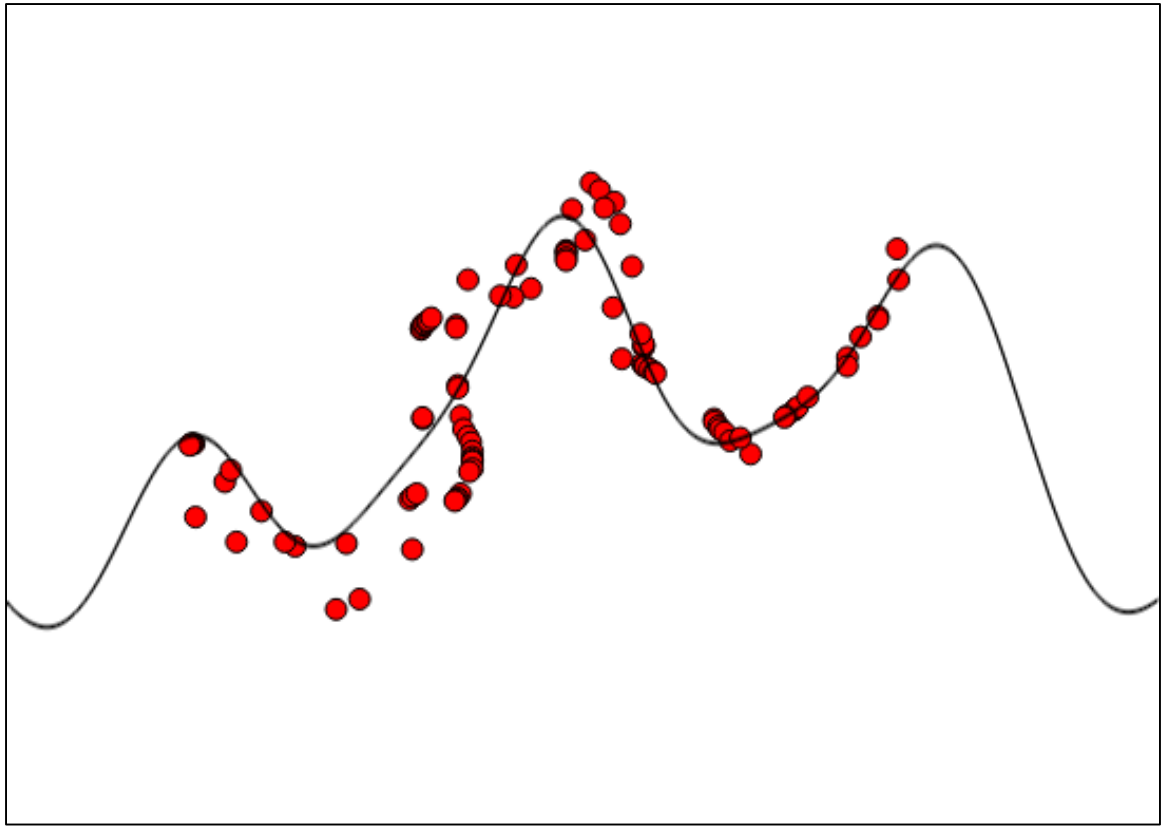
The most frequent item is not necessarily the most relevant one.

Pick data that are relevant and insightful.

Properly choosing the algorithm



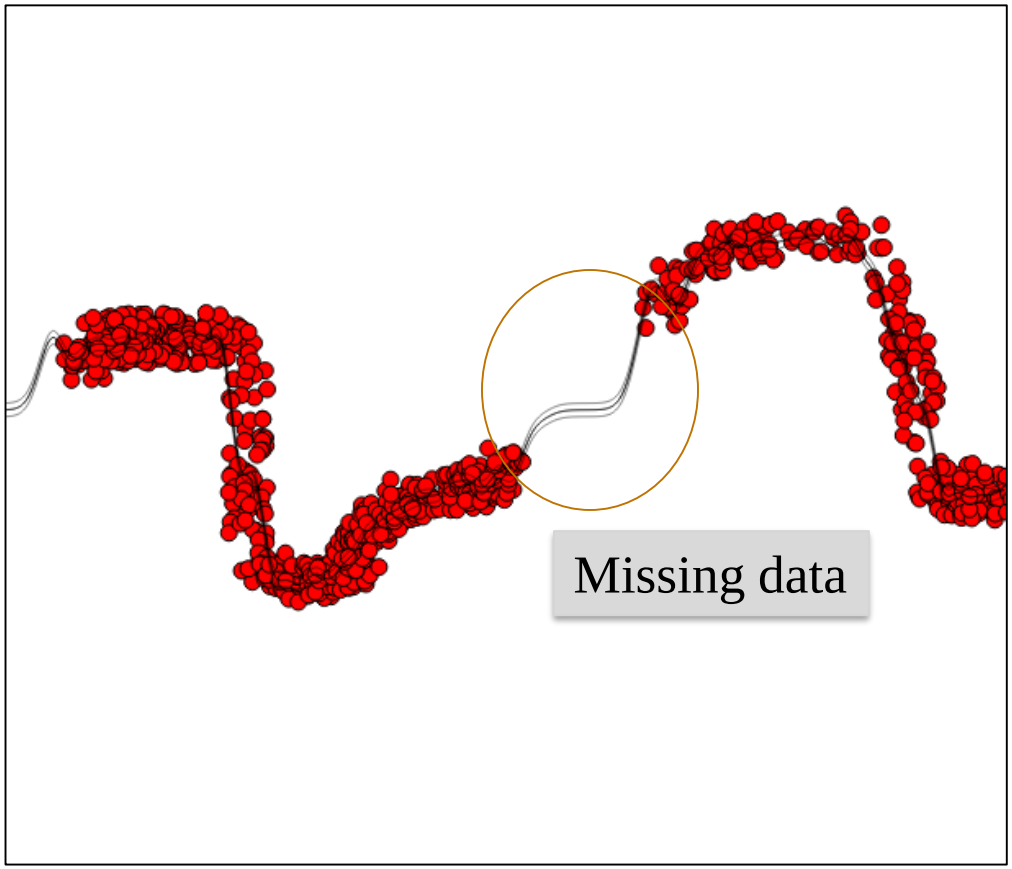
Properly choosing the algorithm



The more complex the model, the more parameters.

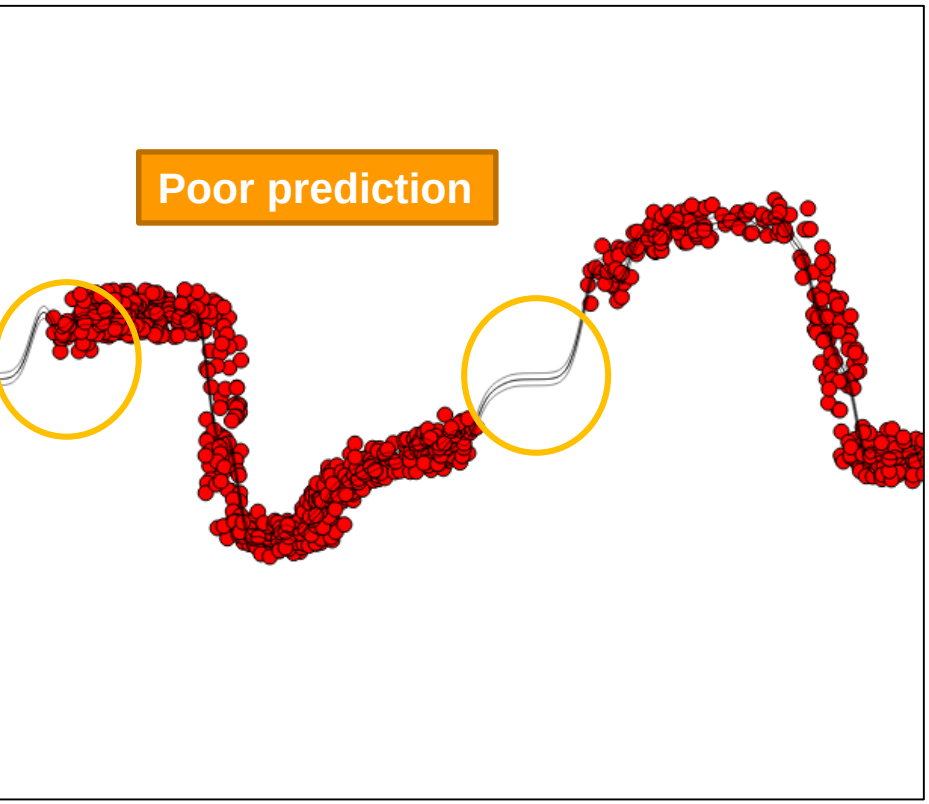
The more parameters, the more data needed to estimate the parameters

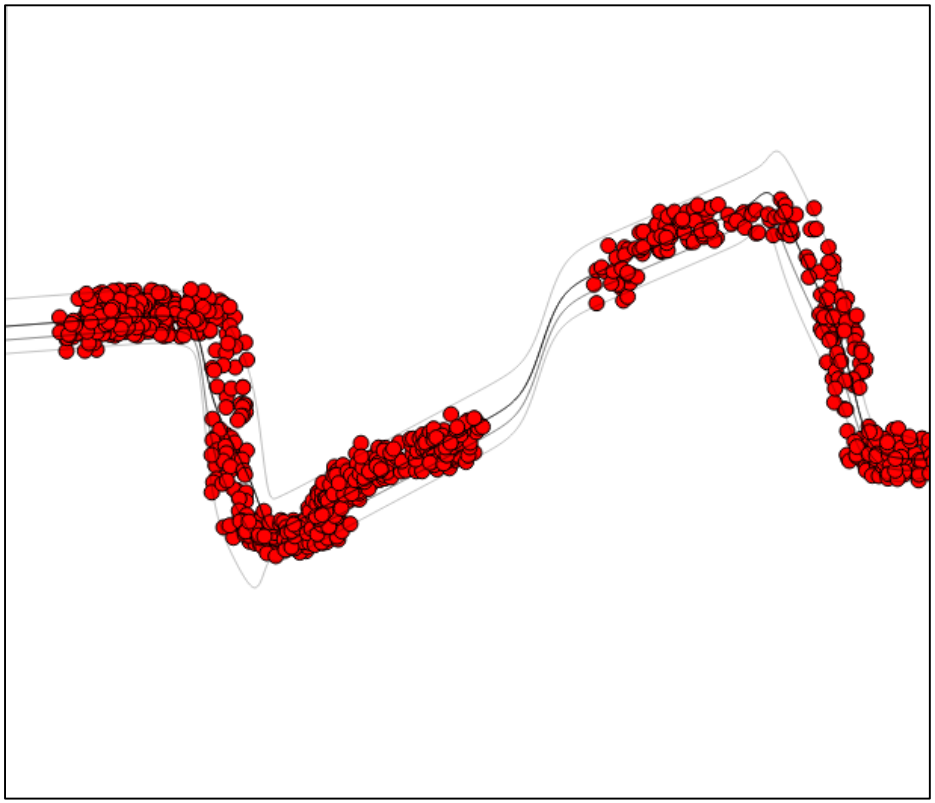
Properly choosing the algorithm



Algorithms differ in sensitivity to missing data

Properly choosing the algorithm



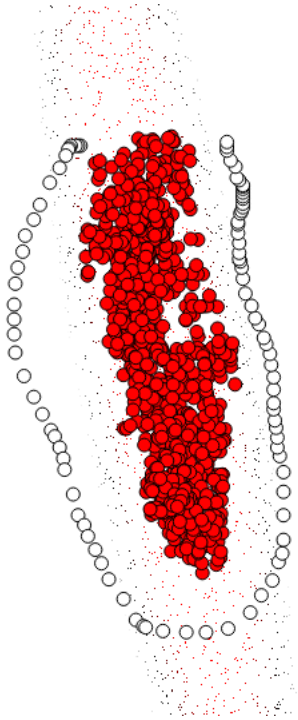


Algorithms differ in sensitivity to missing data

Sensitivity to unbalanced distribution across data

Unbalanced datasets are such that you have much more instances (order of magnitude) of one category than the other: Tumor vs. no tumor

Unbalanced distribution can lead to poor performance when classifying data. The data with least number of instances may be poorly classified.

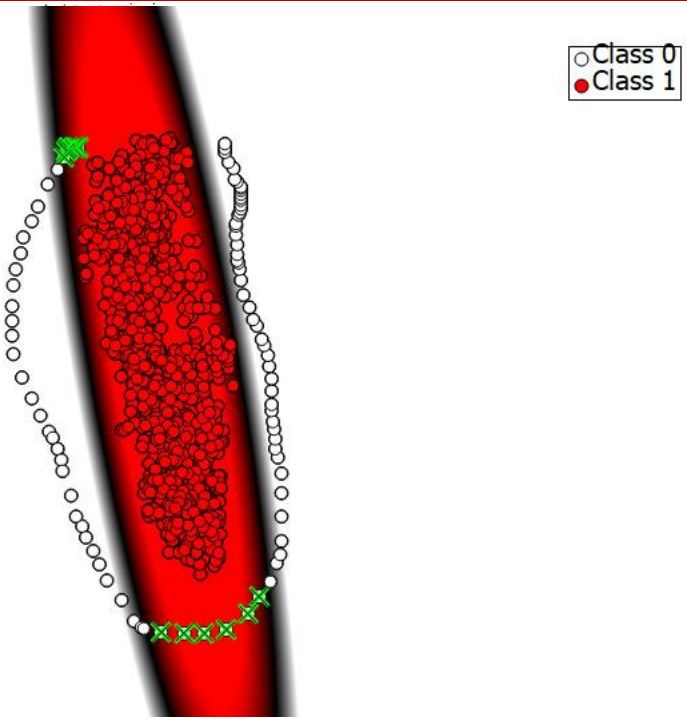


Class 0: 94 instances
Class 1: 920 instance

Sensitivity to unbalanced distribution across data

Unbalanced datasets are such that you have much more instances (order of magniture) of one category than the other: Tumor vs. no tumor

Overall performance good: 98% datapoints correctly classified.
But data from the class with least datapoints is poorly represented.



How to assess performance

There exist many metrics and methods to help you determining how good your choice of data and model is.

Crossvalidation
checks that good performance is not the result of a lucky choice of dataset.

F-measure, Precision, Recall, Accuracy,
determines if your model provides good performance on the data overall

ROC-curve, Confusion Matrix
contrasts performance across classes of data and enables you to determine if performance are good on the data you care most.

Learning Objectives for this Class

At the end of this course, you will have gained:

- ❑ Theoretical knowledge
  - Familiar with the major algorithms in the field
  - Capable of deciding which algorithm to use when, depending on different factors – speed of computation, amount of data required
- ❑ Practical knowledge
  - Familiar with standard procedures and metrics to assess performance
  - Capable of choosing properly a dataset to avoid usual pitfalls

Quiz

Go to moodle and test your understanding of
these two lectures through a quiz.

Take-Home Message

Machine Learning depends on the right choice
for the algorithm as much as for the data.

When you use a machine learning algorithm, you must:

- ❑ Evaluate sensitivity to choice of data through various metrics.
- ❑ Make tough choice on which data to incorporate depending on your "computation budget" at training time and testing time.