Machine Learning

4. Practical tips

Yannick Le Cacheux

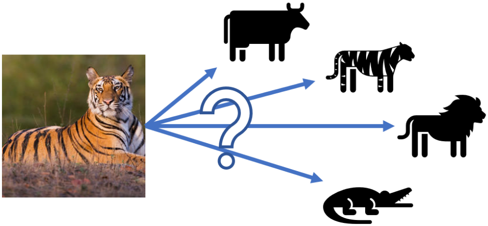
CentraleSupélec - Université Paris Saclay

September 2024

Outline

Multi-class classification

- So far, we have only seen binary classification methods, with two classes or outcomes (for instance dead / alive, will buy / will not buy. . .)
- In practice, we often want to be able to make predictions for more than 1 category



- How can we achieve this?

One-versus-rest classification

- The one-vs-rest or one-vs-all strategy consists in training one binary classifier per class
- The corresponding class is treated as “1”, all the other classes are treated as the same class “0”

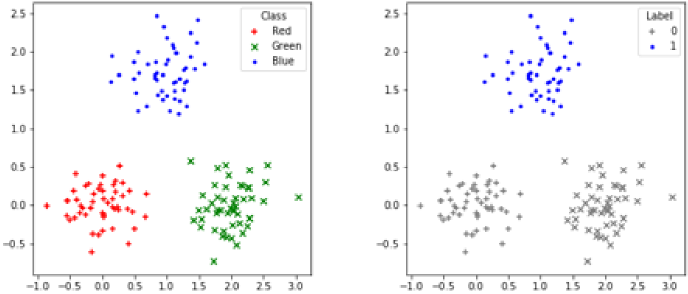


Figure: Example for class blue

One-versus-one classification

- In the one-vs-one strategy, a classifier is trained for each pair of class
- At test time, we can predict the class which received the most "votes" from all classifiers
- While this requires training more classifiers, there are fewer training samples per classifier
- Both one-vs-rest and one-vs-one can be used with any binary classification algorithm

One-versus-one classification

- In the one-vs-one strategy, a classifier is trained for each pair of class
- At test time, we can predict the class which received the most "votes" from all classifiers
- While this requires training more classifiers, there are fewer training samples per classifier
- Both one-vs-rest and one-vs-one can be used with any binary classification algorithm

Multiclass classification

- In some cases, machine learning algorithms can be adapted to directly handle multi-class classification with K output classes
- For instance, with logistic regression: instead of predicting one output such that $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, we can predict one output per class k:

$$f_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x}$$

- To obtain probabilities from f_k, we can use the *softmax* function:

Estimated probability that $\mathbf{x}$ belongs to class k : $\quad \hat{y}_k = \dfrac{\exp(f_k(\mathbf{x}))}{\sum_i \exp(f_i(\mathbf{x}))}$

- We can show that for two classes, we obtain the same estimated probabilities as with the sigmoid function

Multiclass classification

- We also need to adapt the training loss
- The logistic loss can be generalized to the cross-entropy loss:

$$-\left(y\log(\hat{y}) + (1-y)\log(1-\hat{y})\right)$$

$$-\sum_{k=1}^{K} \mathbb{1}[y=k]\log(\hat{y}_k)$$

- The two loss functions are the same with $K = 2$
- We will meet the *softmax* function and the cross-entropy loss again in the deep learning class

Classification and regression

- We have now seen more classification then regression algorithms
- However, some classification methods can easily be adapted to handle regression
- For instance with the decision tree:
  - Instead of measuring uncertainty to decide where to split, we can measure variance
  - Instead of predicting the most common class from the leaf, we can predict the mean value

Outline

Cross-validation

- You should know by now that a dataset should be divided into three parts: a training set, a validation set and a test set
- The models parameters are trained on the training set, the hyper-parameters are selected using the validation set and the final performance is measured on the test set
- However, if we have 100 total samples, each set will not contain many samples

Cross-validation

- We can use k-fold cross-validation: we divide the dataset into k parts, use k-1 parts for training and one for validation or testing, k times.
- The final performance is the average performance.

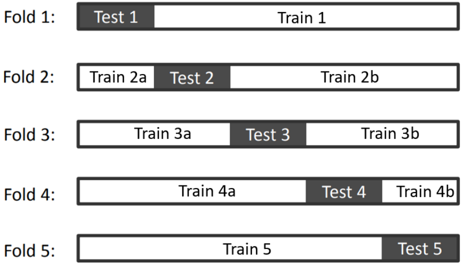


Figure: Example with 5 folds

- When k=N, this is known as leave-one-out cross-validation

Outline

Categorical features

- So far, we mostly considered that input features are continuous values: for instance age, salary, size. . .
- What happens if we have categorical input variables, for instance gender, eye color or town?
- We could represent the eye colors blue, brown and green by the values 1, 2 and 3.

(a) 1 (b) 2 (c) 3

Figure: Discrete numbers for categories

- Does it make sense?

Feature engineering

- It is usually a better solution to use *one-hot encoding*:

Eye color
Blue
Brown
Brown
Green
Blue
Brown

Eye color blue	Eye color brown	Eye color green
1	0	0
0	1	0
0	1	0
0	0	1
1	0	0
0	1	0

Table: Left: categories, right: one-hot encoding

- Each categorical feature is replaced by as many binary features as there are categories

Feature engineering

- In some cases, it can still make sense to use non binary numbers for categories (for instance with cloth sizes S, M, L, XL...)
- Sometimes there are many (thousands) categories: we can keep a fixed number of categories, and label the other as a single category "other"
- Many questions: how many categories do we keep?
  - Some categories with few samples may have a large impact on the result
  - Some categories with many samples may have no impact
- Feature engineering may be something of an art form.

Feature engineering bis

- Suppose you want to predict the price of the house. One of the input features is the construction year of the house. Is it a useful feature?
- Could we create a more meaningful feature to represent this information?
- Feature engineering is usually responsible for a very large part of the performance of a model.

Outline

Missing values

- In practice, some values can be missing for some samples and some features

Surface area (m^2)	Distance (km)	Price (€)
62	3	631,000
128	-	1,150,000
12	2	152,000
-	-	370,000
55	3	540,000

- What can we do in this case?

Missing values

- Some possibilities
  - Discard corresponding training samples
  - Replace by the mean of the missing feature (or the most frequent category)
  - Try to predict the missing value from the other features of the corresponding sample
- However, be careful! Sometimes missing values are not random, and treating them as such may bias the model
-   - Always question the process by which the data was obtained

Weird values

- Similarly, some values may seem out of place (outliers)

Area (m^2)	Distance (km)	Price (€)
62	3	6,310,000,000
1280	8	1,150,000
12	2	152,000
35	124	180,000
55	3	540,000

Table: One value is in square feet and not in square meter
One value is rubbish and can be discarded
One value is unusual but true

- There are again several possibilities, but you should always be careful about their implications.