Evaluation of machine learning results

Introduction

How do we evaluate the performance of an ML model?

How good is it at predicting? How well it has learned?

1. Regression
Mean Squared error (MSE)

We have seen some evaluation for linear regression with

- the squared error as a loss function and a measure of the performance
- the R-squared value (see previous slides on linear regression)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

True label

Predicted value

Mean Average Error (MAE)

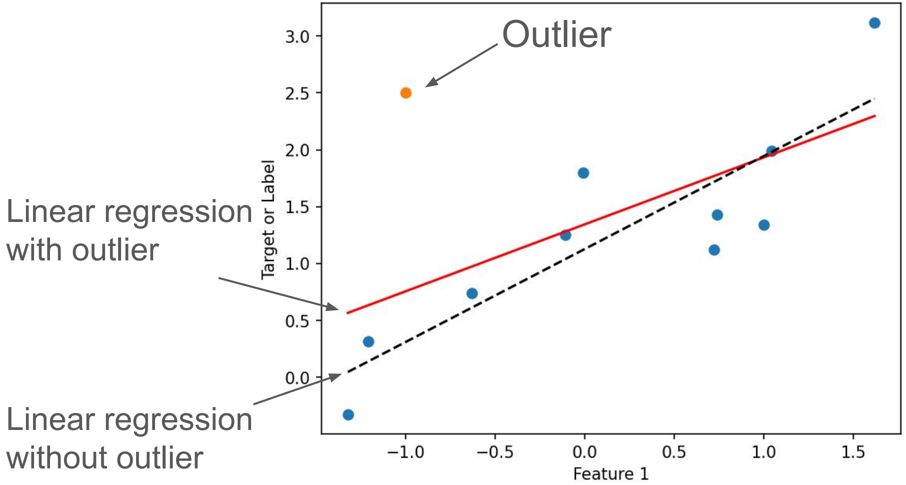
What can go wrong with Mean Squared error?

- The error can become enormous with the square
- It can be a problem when there are outliers, or mislabelled samples

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Better solution:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

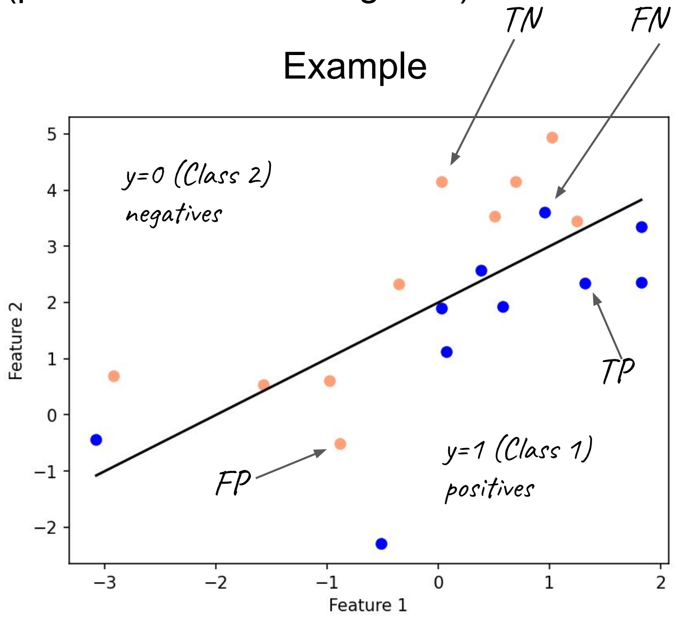


2. Classification
Confusion matrix

- TP: True positive (positive classified positive)
- TN: True negative (negative classified negative)
- FP: False positive (negative classified positive)
- FN: False negative (positive classified negative)

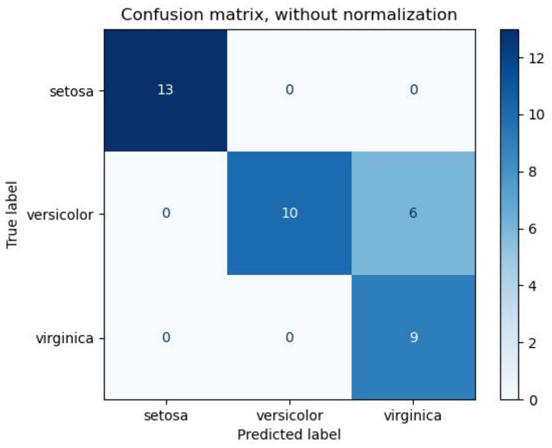
Predicted label

True label		1	0
	1	TP	FN
	0	FP	TN



Confusion matrix

Example for more than 2 classes:



From
https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html

Accuracy

Summarize in one number:
Accuracy = number of correct predictions / total number of predictions

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Predicted label

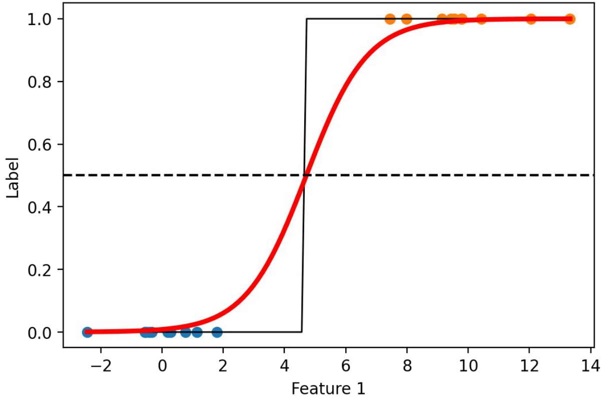
True label

	1	0
1	TP	FN
0	FP	TN

AUC and ROC curve

- Area Under the Curve
- Receiver Operating Characteristic

Give an overview of the influence of the class threshold



Threshold

Class 1,
Negatives

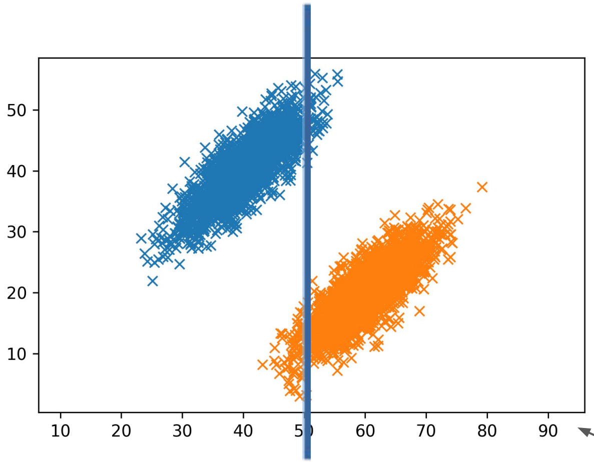
Class 2,
Positives

Threshold

Threshold 0.5 seems a good number,
but maybe there is a better one…

We don’t want to depend on this threshold for evaluating the model

- Area Under the Curve
- Receiver Operating Characteristic

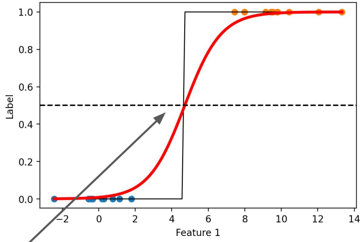


Class 1,
Negatives

Class 2,
Positives

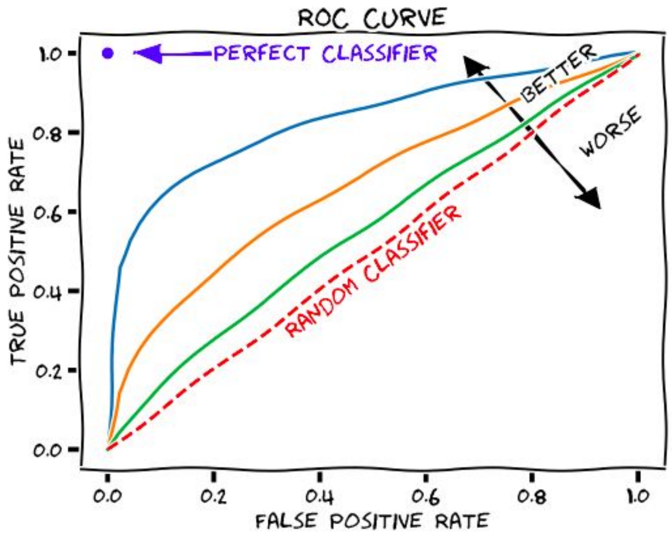
Threshold

Each threshold value gives a point of the curve



Threshold

Is this feature good for classifying this dataset?



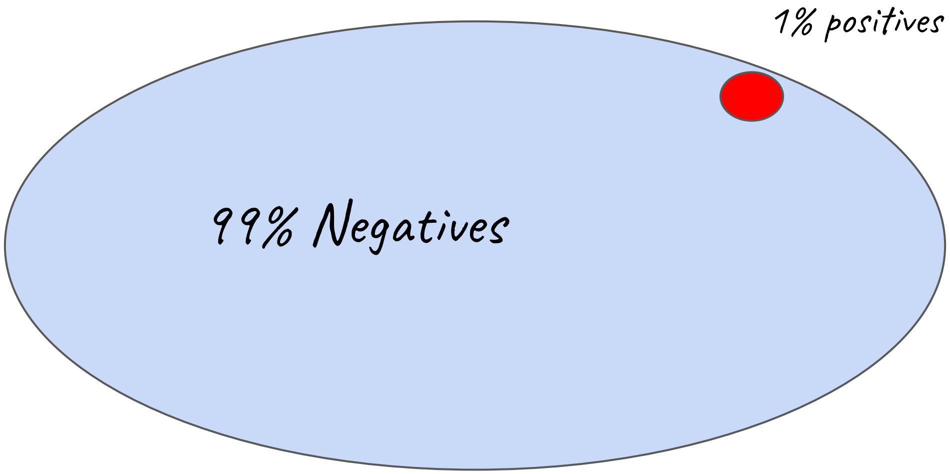
MartinThoma, CC0, public domain, via Wikimedia Commons

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$

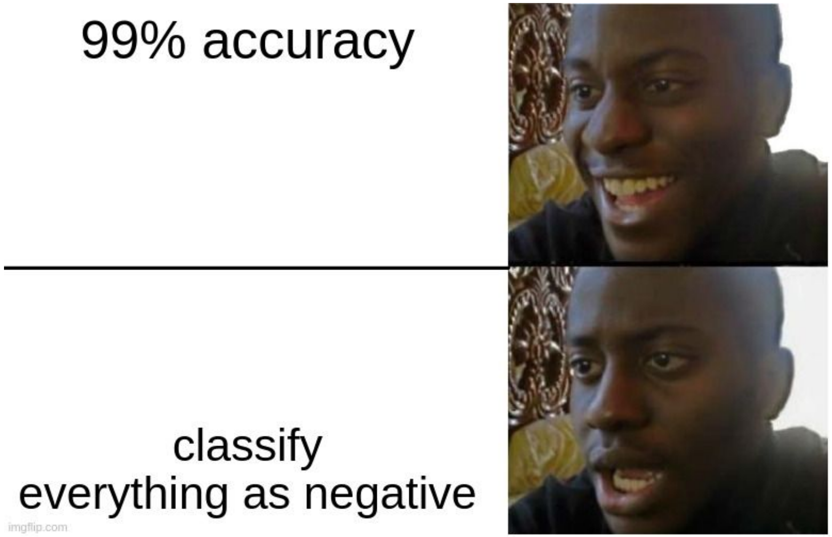
$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = 1 - \mathrm{TNR}$$

AUC: a single number

Class imbalance



Classify everything as negative: Accuracy = 0.99 !



Measures when class distinction is important

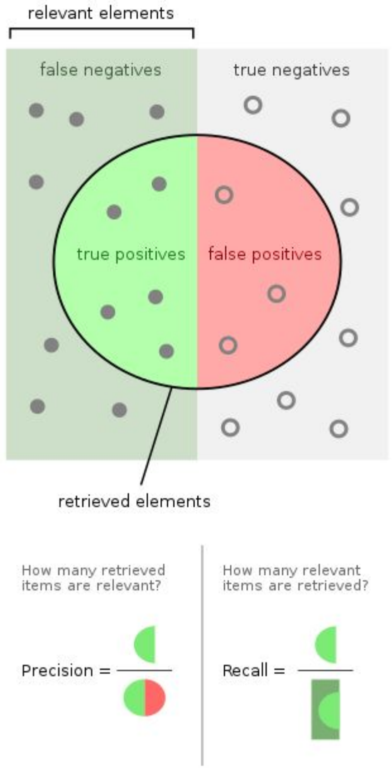
<u>Example in medicine:</u> Healthy (negative) / non-healthy (positive)
It may be better to have false positives than false negatives

Table 8.2 Evaluation Measures

Term	Definition	Calculation
Sensitivity	Ability to select what needs to be selected	TP/(TP + FN)
Specificity	Ability to reject what needs to be rejected	TN/(TN + FP)
Precision	Proportion of cases found that were relevant	TP/(TP + FP)
Recall	Proportion of all relevant cases that were found	TP/(TP + FN)
Accuracy	Aggregate measure of classifier performance	(TP + TN)/ (TP + TN + FP + FN)

TP, *true positive;* FP, *false positive;* FN, *false negative;* TN, *true negative.*

From: Data Science: Concepts and practice,
Vijay Kotu, Bala Deshpande



From Walber, CC BY-SA 4.0, via Wikimedia Commons

F1-score

A score to account for imbalanced classes (with small amount of positives)

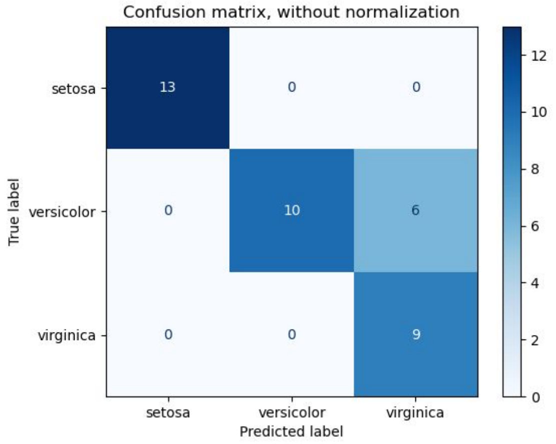
$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{tp}}{2\mathrm{tp} + \mathrm{fp} + \mathrm{fn}}$$

Intuitively,

- precision is the ability of the classifier not to label as positive a sample that is negative,
- recall is the ability of the classifier to find all the positive samples,
- F1-score is the “harmonic” mean of them (if one of them is low, F1 is low)

Multiclasses

How to deal with more than 2 classes?



- Take one class versus the rest. This gives a score per class.

If one single value is needed:

- Macro-averaged score: average the score of the different classes (from “one vs the rest” scores)
- Micro-averaged score: compute globally the number of TP, FP and FN (FP = FN in this case, as FP for a class is a FN for another class). Precision, recall, accuracy and F1 are equal!

Micro average better if class imbalance