MachineLearnAthon - Microlecture
Evaluating Classification Models

Classification

15.10.2024

Recap previous microlectures Classification-1,2

- Types of Machine Learning Tasks
- Stages of Supervised Machine Learning Pipeline
- Proper Machine Learning Modeling
- Linear Classification Models
- Non-Linear Classification Models

Learning outcomes of today

After successfully completing this micro-lecture, you are able to….

- Choose the appropriate evaluation metric for the classification task
- Understand the probabilistic classification
- Measure the classification model reliability
- Understand what is classifier calibration

Agenda for today

- Evaluation of Classification Models
  - Confusion Matrix
- Probabilistic Classification
  - Measuring model reliability
  - Calibration

How to evaluate your classification model?

- **Classification Example:** (A model that diagnose COVID-19)
  - Infected Person = Positive Class
  - Healthy Person = Negative Class

- A sample of 100 persons (2 infected and 98 healthy)

- Model Predictions Table

Predicted Positive (Out of 2)	Predicted Negative (Out of 98)	Accuracy	Recall	Precision	F1-Score
0	98	98%	0%	Undef	0%
2	96	93%	95%	50%	66.7%
2	0	2%	100%	2%	4%

How to evaluate your classification model?

- Accuracy and Error Rate are the primary evaluation metrics of classifiers.

  - **Accuracy** = Proportion of correctly classified instances
  - **Error Rate** = 1 - Accuracy

$$acc = \frac{1}{|Te|} \sum_{x \in Te} I[\hat{c}(x) = c(x)]$$

- I[.] equals to 1 if the argument . evaluates to true and 0 otherwise.
- c(x) = ground truth of instance x, c^(x) = prediction of instance x

How to evaluate your classification model?

- Accuracy is usually not important. WHY?

- Back to Salmon (+ve class) Vs Sea Bass (-ve class) Classification Task:
  - The cost of misclassifying salmon as sea bass *(False Negative)* is that the end customer will be …. if he finds occasionally find a tasty piece of salmon when he purchases sea bass.

  - The cost of misclassifying sea bass as salmon *(False Positive)* is …. when he finds a piece of sea bass purchased at the price of salmon.

How to evaluate your classification model?

It is often useful to see the kind of errors that the classifier makes (Confusion Matrix)

Example:

	Predicted +ve	Predicted -ve	
Actual +ve	True Positives	False Negatives	
Actual -ve	False Positives	True Negatives	
			Total

	Predicted +ve	Predicted -ve	
Actual +ve	30	20	50
Actual -ve	10	40	50
	40	60	100

How to evaluate your classification model?

- Accuracy = (TP + TN) / Total
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)

	Predicted +ve	Predicted -ve	
Actual +ve	True Positives	False Negatives	
Actual -ve	False Positives	True Negatives	
			Total

- Back to Salmon (+ve class) Vs Sea Bass (-ve class) Classification Task:
  - Misclassifying salmon as sea bass *(False Negative)*
  - Misclassifying sea bass as salmon *(False Positive)*

Accuracy = 88%
Precision = 83%
Recall = 50%

	Pred +ve	Pred -ve	
Actual +ve	10	10	20
Actual -ve	2	78	80
	12	88	100

Accuracy = 90%
Precision = 50%
Recall = 100%

	Pred +ve	Pred -ve	
Actual +ve	20	0	20
Actual -ve	20	60	80
	40	60	100

Probabilistic Classification

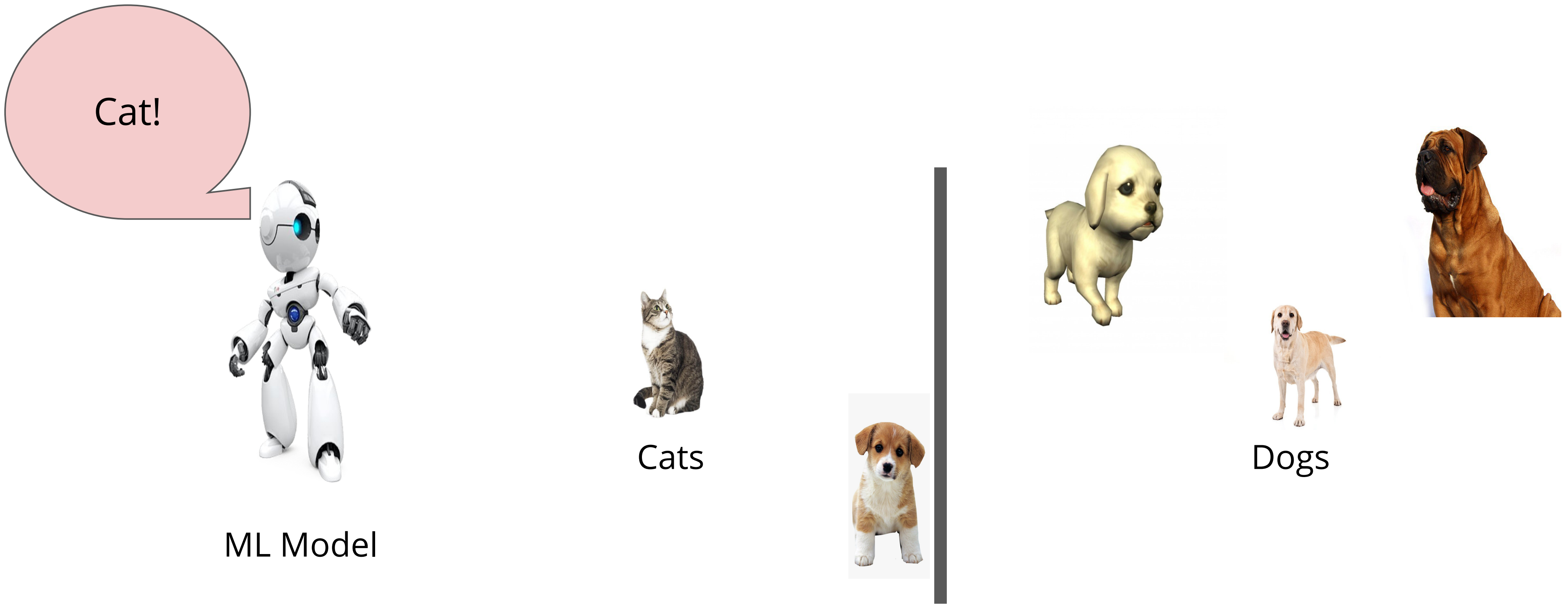
Input Image

ML Model
99% accurate

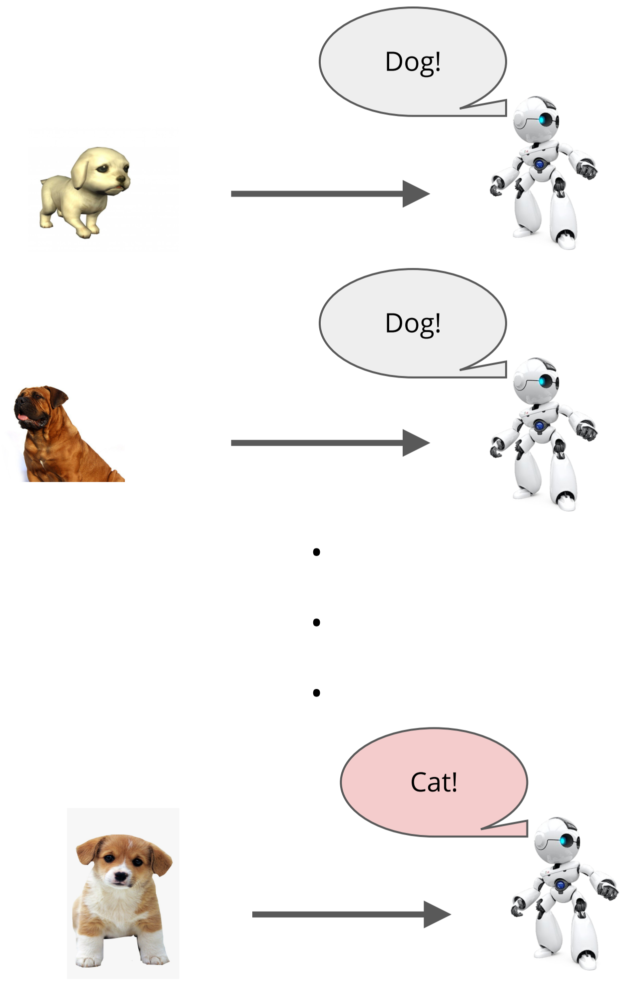
Cats

Dogs

Probabilistic Classification

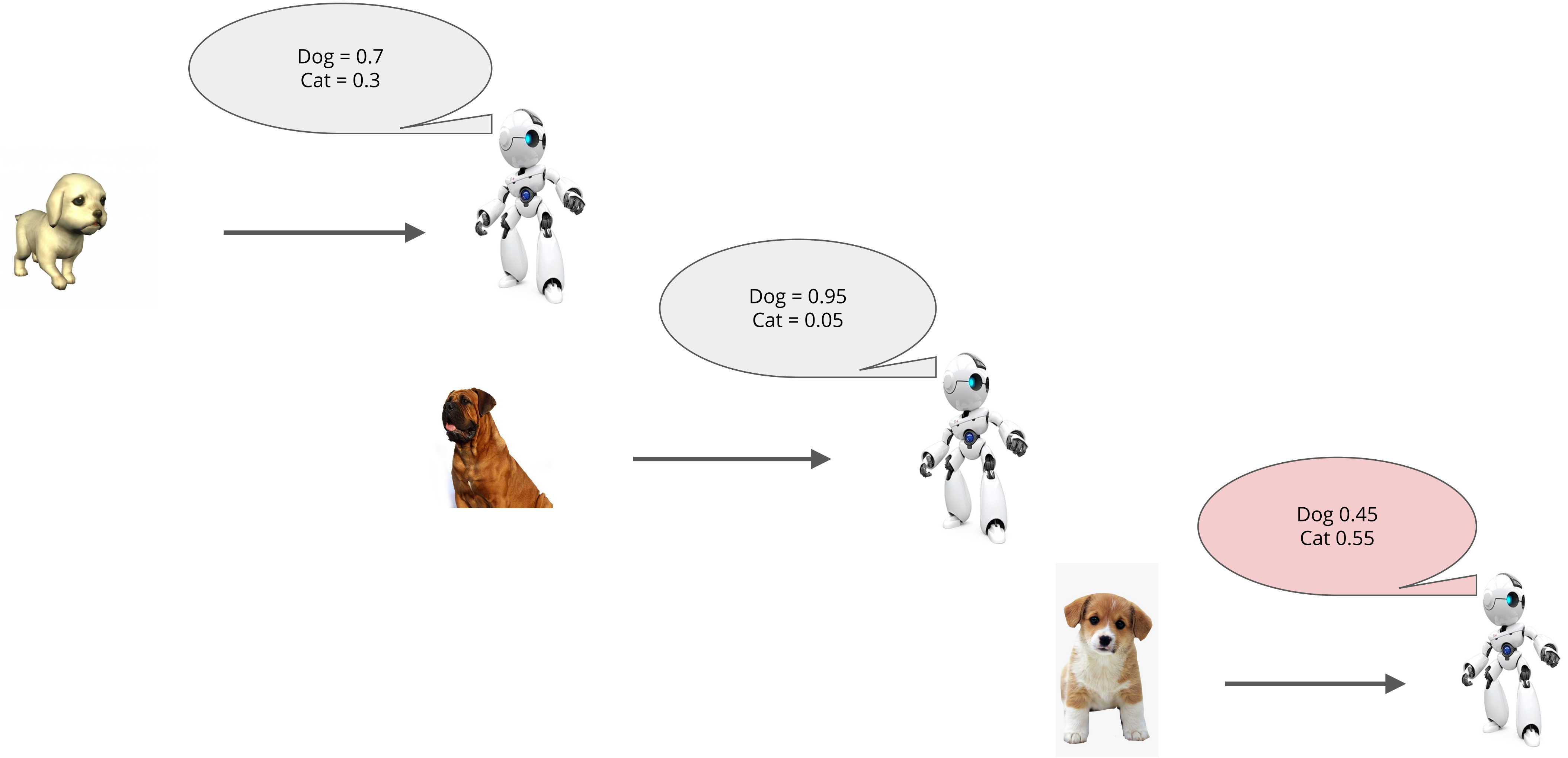


Probabilistic Classification



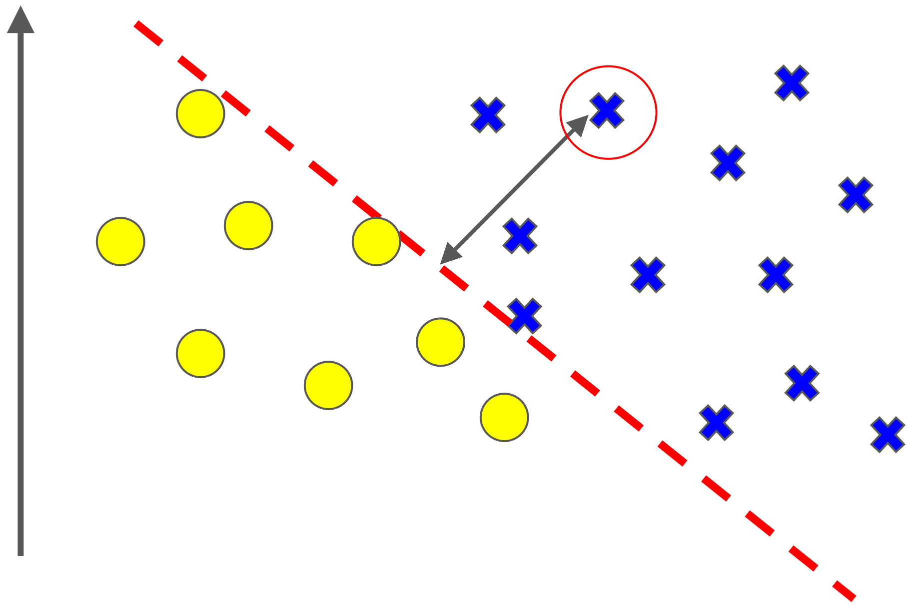
- Still 99% Accurate.
- In sensitive domains, I can not tolerate many wrong decisions especially when data distribution changed.
- Probabilistic Classification is needed.

Probabilistic Classification



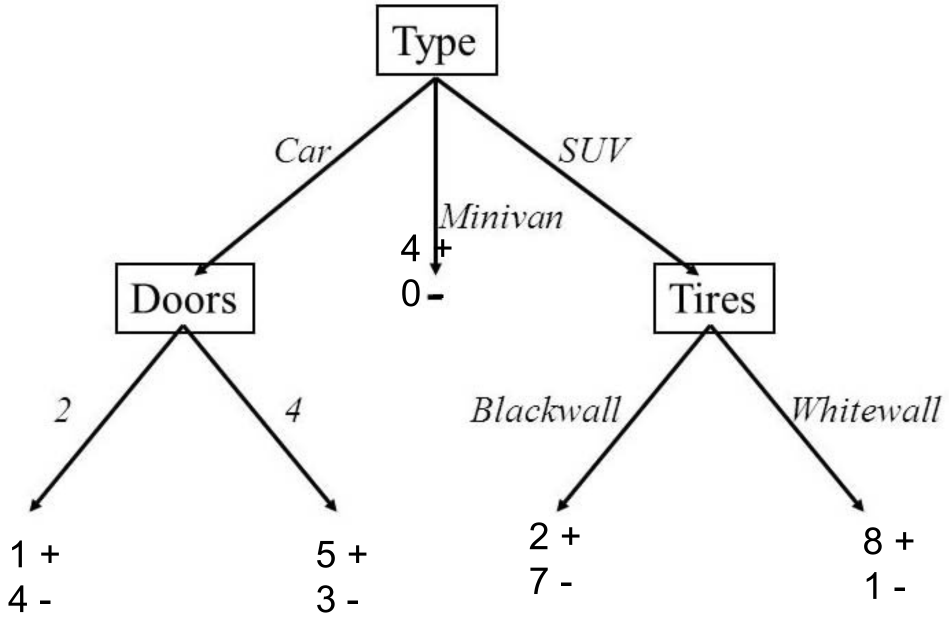
Probabilistic Linear Classifier

- Can the model predict probabilities instead of labels only?
  - Linear models: distance to the decision boundary



Probabilistic Decision Tree

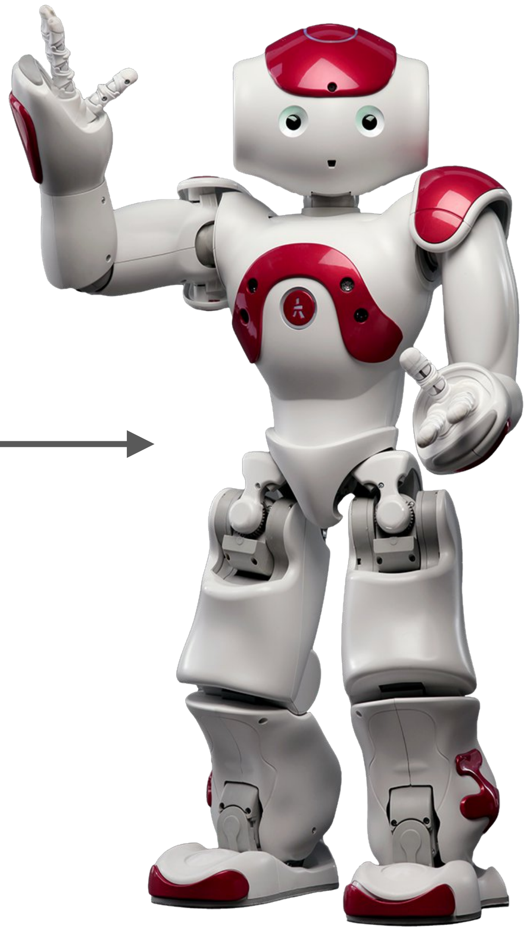
- Can the model predict probabilities instead of labels only?
  - Decision Tree



SUV with Blackwall tires has probability 2/(2+7) = 0.22 to be a taxi!

How to measure Calibration? Reliability Diagram

Input

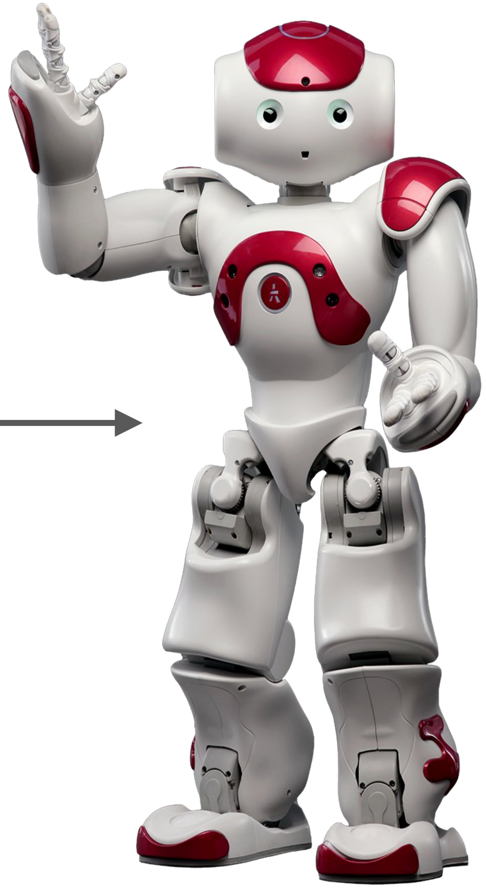


1000 patients

How to measure Calibration? Reliability Diagram

Input

1000 patients



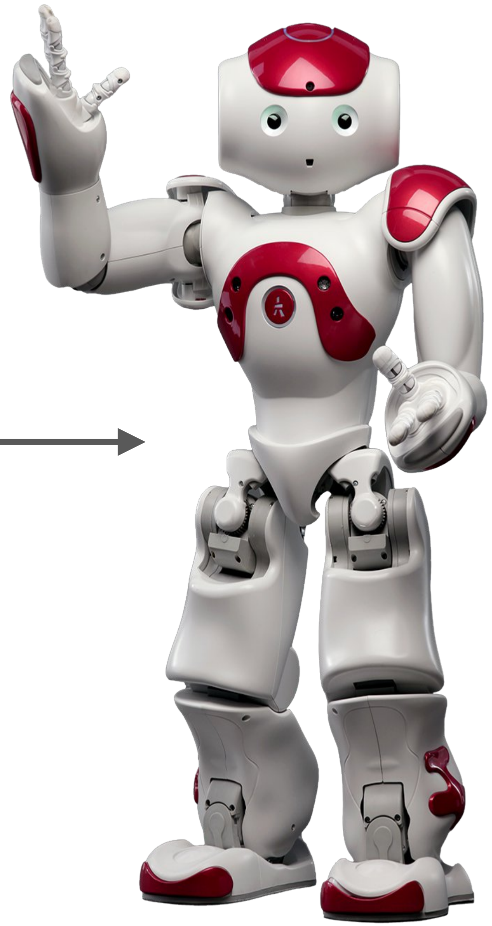
- 90 patients with predicted probability from 0 to 0.1 (average 0.05)
- 2% of those patients were actually Covid patients

How to measure Calibration? Reliability Diagram

0.05 → 2% (90)

Input

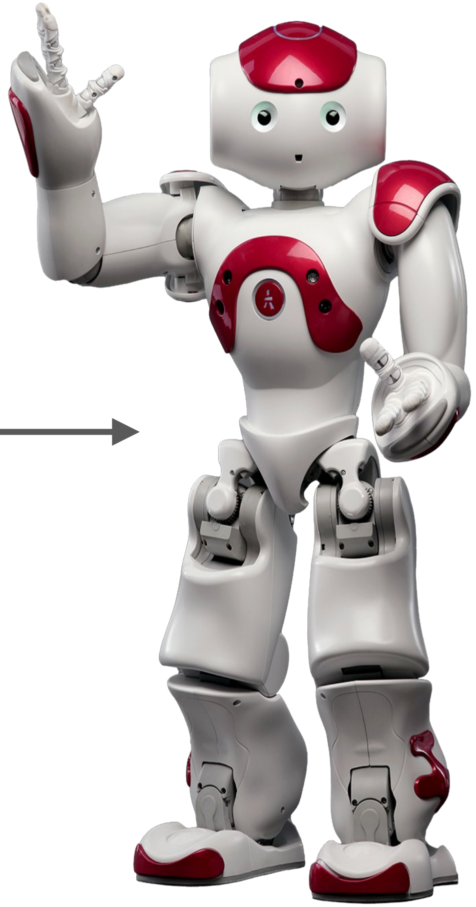
1000 patients



How to measure Calibration? Reliability Diagram

Input

1000 patients



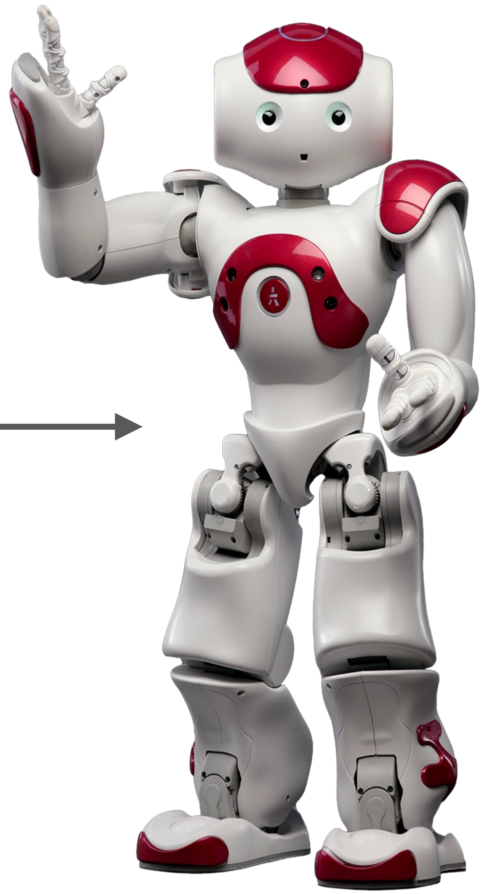
0.05 → 2% (90)

- 110 patients with predicted probability from 0.1 to 0.2 (average 0.17)
- 13% of those patients were actually Covid patients

How to measure Calibration? Reliability Diagram

Input

1000 patients



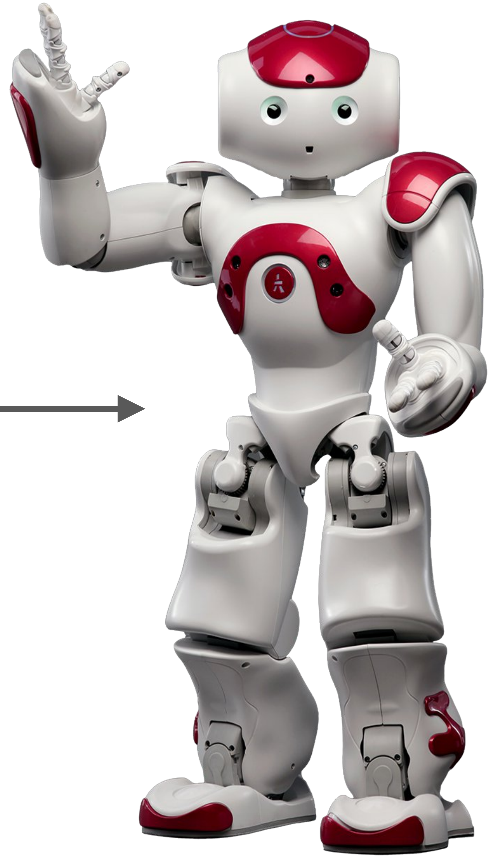
0.05 → 2% (90)

0.17 → 13% (110)

How to measure Calibration? Reliability Diagram

Input

1000 patients



0.05 → 2% (90)

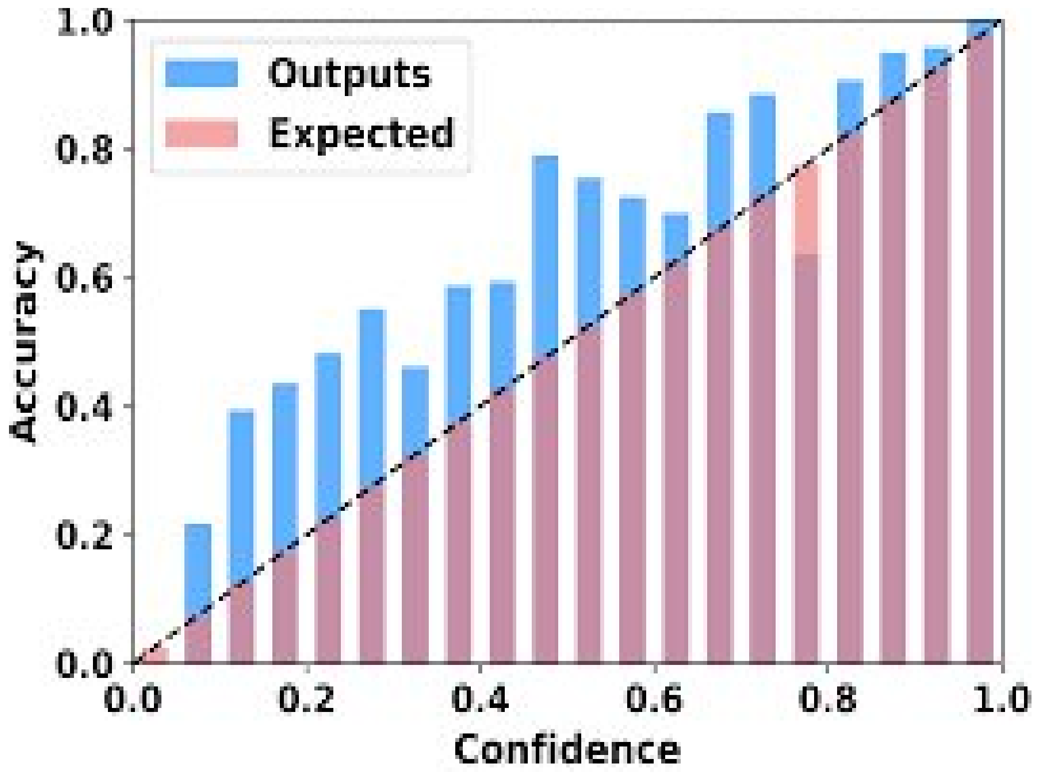
0.17 → 13% (110)

•
•
•

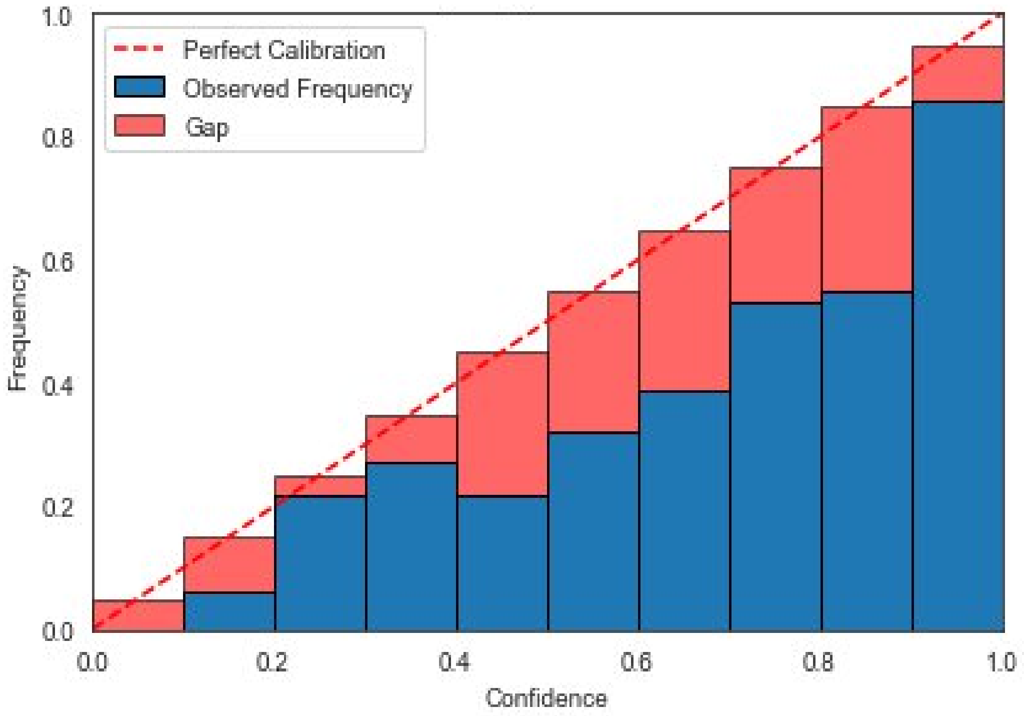
0.98 → 92% (290)

How to measure Calibration? Reliability Diagram

Overconfident Uncalibrated Model



Underconfident Uncalibrated Model

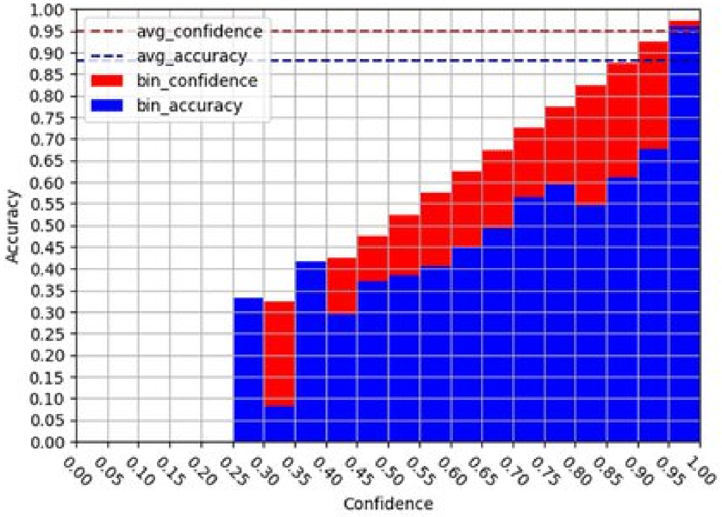


How to measure Calibration? Reliability Diagram

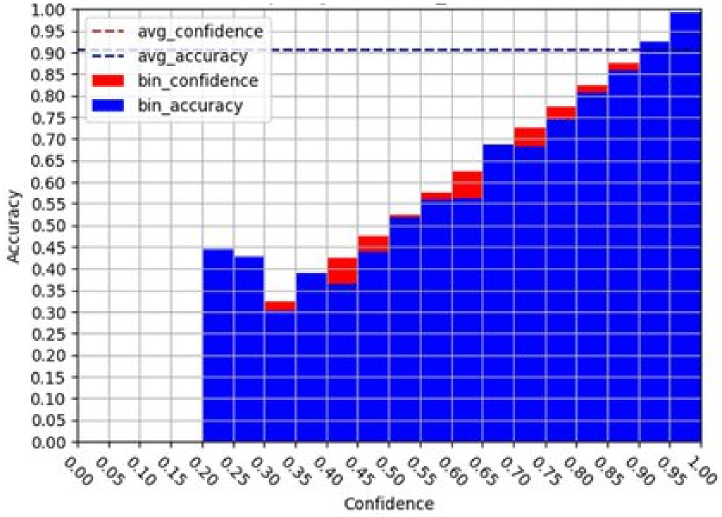
Calibration of a classifier

A post-processing step where we take a trained model to improve its predicted confidence levels to match the accuracy of these predictions.

Before



After



Recap this lecture

After successfully completing this lecture, you are able to….

- Choose the appropriate evaluation metric for the classification task
- Understand the probabilistic classification
- Measure the classification model reliability
- Understand what is classifier calibration

Outlook: What will the tutorial be about?

Your model gives you an accuracy of 85%, but can you really trust it? Is it making overly confident or overly cautious predictions?

- In this micro-lecture tutorial, we will uncover the secrets behind proper model evaluation—beyond just accuracy. You'll learn how metrics like:
  - precision, recall, F1-score, and ROC-AUC help give a more complete picture of performance.
- Additionally, we'll dive into calibrating classifiers, ensuring that your probability estimates are as reliable as they seem.
- By the end, you'll be equipped with the knowledge to build not just accurate, but trustworthy classification models!

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.