Introduction to Machine Learning

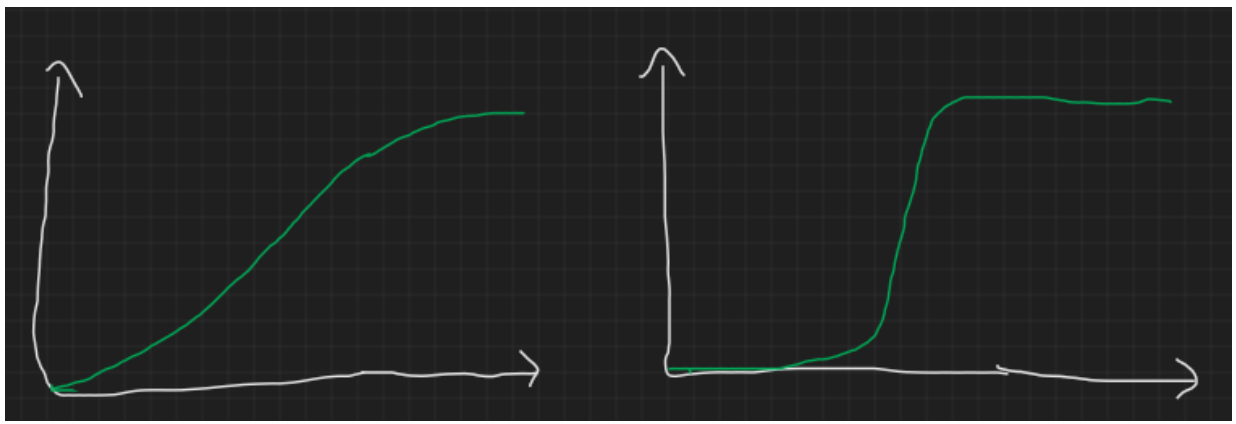
Evaluation: Discrimination & Calibration

Learning goals

- Understand the concepts of discrimination and calibration
- Understand that they are sometimes at odds

DISCRIMINATION

- Consider, again, the binary classification case.
- **Discrimination** is the ability of a classifier to perfectly separate the population into positive and negative instances.
  - The classifier is said to discriminate well if predictions differ strongly across classes – e.g., predicted probabilities for the negative (positive) class are all close to zero (one).
  - Measures of discrimination: e.g., AUC, sensitivity, specificity.



CALIBRATION

- **Calibration**, on the other hand, assesses the concordance of predicted probabilities with the observed outcome (for any reasonable grouping).
  $\rightarrow$ For scoring classifiers, evaluating calibrations requires transformation of scores to posterior probabilities first.
- Predictions of a well-calibrated classifier follow approximately the same distribution as the true data labels.
- Poor calibration occurs with imbalanced classes or when the learner lacks a probabilistic framework (e.g., k-NN, trees).
- We distinguish two different notions of calibration:
  - **Calibration in the large** is a property of the *full* sample.
    $\rightarrow$ Observed class-1 frequency in full sample vs average overall predicted class-1 probability.
  - **Calibration in the small** is a property of *subsets*.
    $\rightarrow$ Observed likelihood in subset vs average predicted class-1 probability in that subset.

CALIBRATION AND DISCRIMINATION

- A well-calibrated classifier can be poorly discriminating.
- E.g., consider two probabilistic classifiers f_1 and f_2:

observation nr.	truth	prediction f_1	prediction f_2
1	1	0.9	0.6
2	1	0.9	0.6
3	1	0.9	0.4
4	0	0.1	0.4
5	0	0.1	0.4
6	0	0.1	0.6
avg. class-1 prob.	50%	50%	50%

- Both classifiers have identical calibration in the large (50%), but clearly, f_1 has better discriminative power.

CALIBRATION AND DISCRIMINATION

- Conversely, a good discriminator can have bad calibration:

observation nr.	truth	prediction f_1	prediction f_2
1	1	0.97	0.99
2	1	0.97	0.99
3	0	0.01	0.67
4	0	0.01	0.67
5	0	0.01	0.67
6	0	0.01	0.67
7	0	0.01	0.67
8	0	0.01	0.67
avg. class-1 prob.	25%	25%	75%

- Both classifiers discriminate well (e.g., setting thresholds at 0.5 and 0.8, respectively).
- Classifier f_2 is, however, rather poorly calibrated: the probability of class 1 would be estimated at three times the true proportion.