CLASSIFICATION

"A or Not A" — Pick prob. threshold and that's your hypothesis.

But need new eval metrics! For linear we used RMSE (measuring dist from prediction to label). For logistic:

- Accuracy? Fraction of predictions correct
  - ↳ Imbalance problems when positive/negative is very rare (in the distribution)
  - Consider {true, false} {pos., neg.}

- Precision and Recall

$\left(\frac{TP}{TP+FP}\right)$ (Predicted true — was it true?) (How many trues did we predict true?) $\left(\frac{TP}{TP+FN}\right)$

These are in tension:
- too much precision can lower recall
- vice versa

ROC Curve (Receiver Operating Characteristic) Evaluate every possible class. threshold

For each threshold $\in [0,1]$, let
x = false pos. rate @ threshold and
y = true pos. rate @ threshold. Then
the ROC curve is defined by $\{x(t), y(t) \quad \forall t \in [0,1]\}$

scale invariant (not comparing abs. values)

AUC (area under ROC curve) $= \int_{t=0}^{t=1} y(t)\, dx(t)$

threshold invariant

(For random pos. + random neg., what is P(model ranks correctly)? — this is AUC)

Prediction bias : want predictions to match observations. Avg.(pred.) = Avg(obs.)

↳ Easy to fool but is a good canary.

↳ Too much bias?? Consider:

- incomplete feature set
- buggy pipeline
- biased training sample
- too much regularization

→ Fix in the model, not "calibration layer"