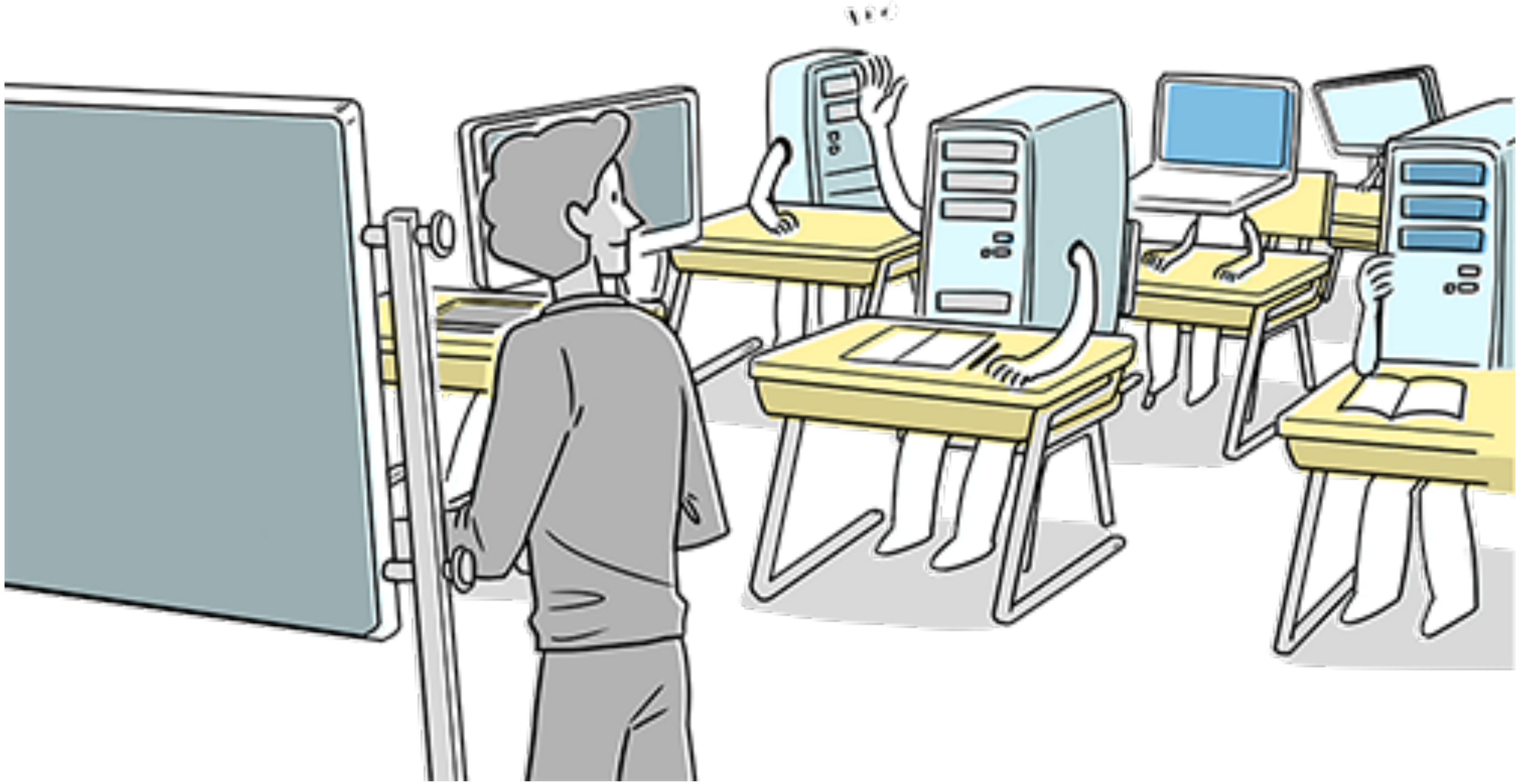


EVALUATION METRICS

Metrics for classification task

- Confusion Matrix
- Accuracy
- Precision/Recall
- ROC Curve and AUC

Confusion matrix

Understanding what types
of mistakes a learned model makes

activity recognition from video

actual class

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	100	0	0	0	0	0	0	0	0	0
jack	0	100	0	0	0	0	0	0	0	0
jump	0	0	89	0	0	0	11	0	0	0
pjump	0	0	0	100	0	0	0	0	0	0
run	0	0	0	0	89	0	11	0	0	0
side	0	0	0	0	0	100	0	0	0	0
skip	0	0	0	0	0	0	100	0	0	0
walk	0	0	0	0	0	0	0	100	0	0
wave1	0	0	0	0	0	0	0	0	67	33
wave2	0	0	0	0	0	0	0	0	0	100

predicted class

Confusion matrix for 2-class problems

		actual class	
		positive	negative
predicted class	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

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

Is accuracy an adequate measure of predictive performance?

Accuracy may not be useful measure in cases where

- There is a large class skew
  - Is 98% accuracy good if 97% of the instances are negative?

- There are differential misclassification costs – say, getting a positive wrong costs more than getting a negative wrong
  - Consider a medical domain in which a false positive results in an extraneous test but a false negative results in a failure to treat a disease

Other accuracy metrics

		actual class	
		positive	negative
predicted class	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

$$\text{recall (TP rate)} = \frac{\text{TP}}{\text{actual pos}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{precision} = \frac{\text{TP}}{\text{predicted pos}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

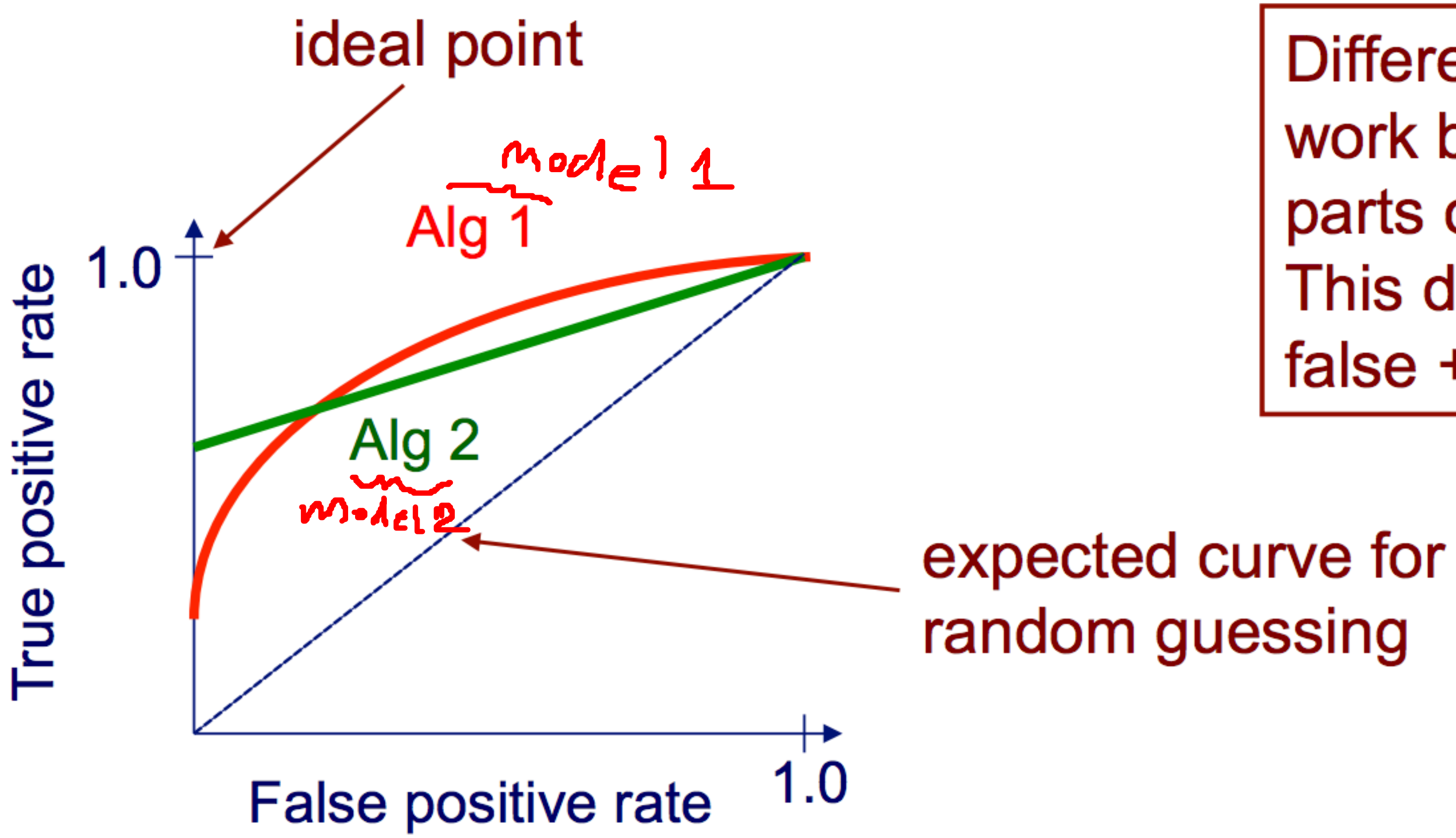
Control the trade-off between precision and recall

Hard to

Cannot optimize
Both recall & precision
at same time

ROC Curve

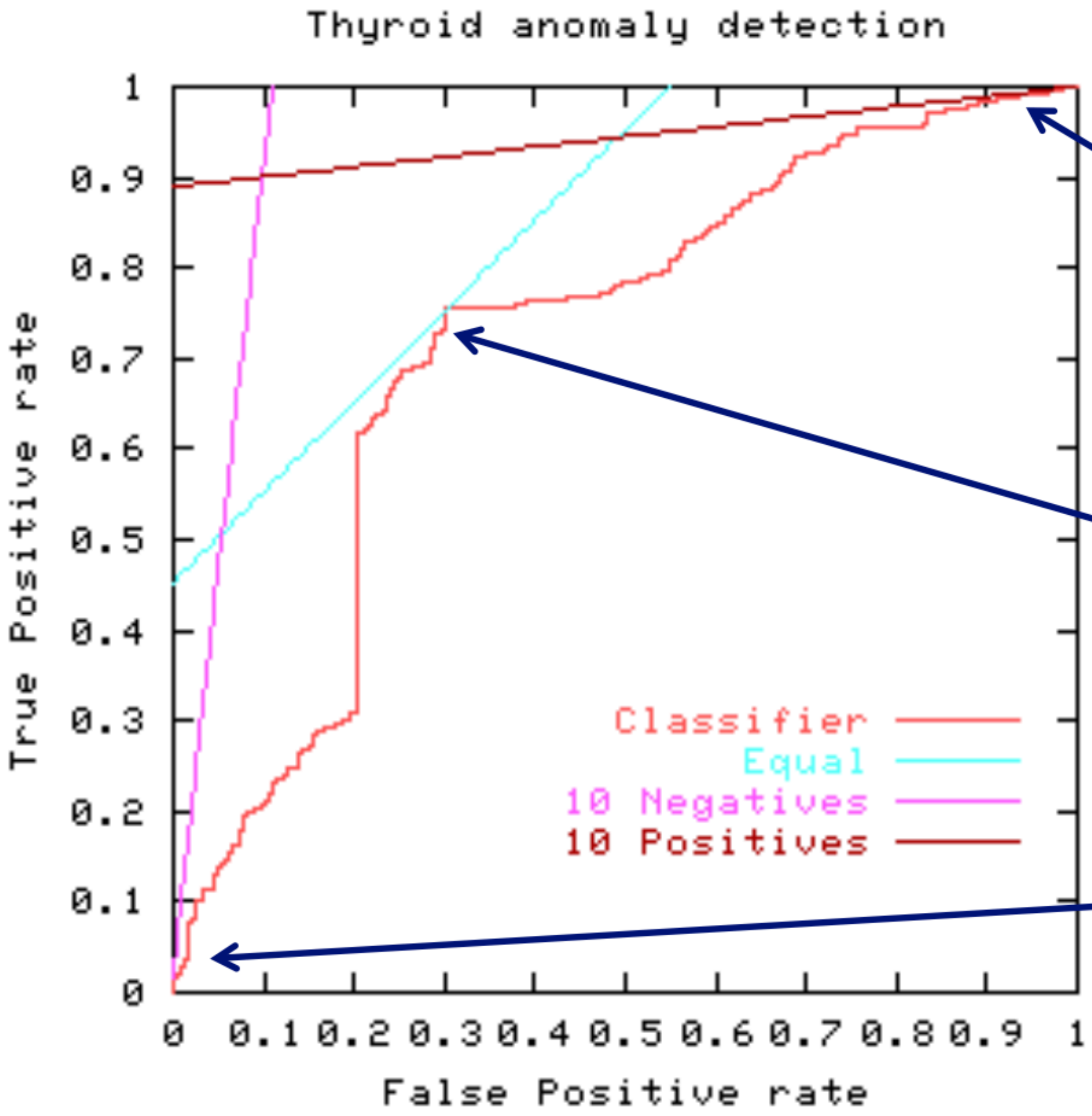
A *Receiver Operating Characteristic* (*ROC*) curve plots the TP-rate vs. the FP-rate as a threshold on the confidence of an instance being positive is varied



Different methods can work better in different parts of ROC space. This depends on cost of false + vs. false -

ROC curves and misclassification costs

Plot and seleted threadhold
By objective

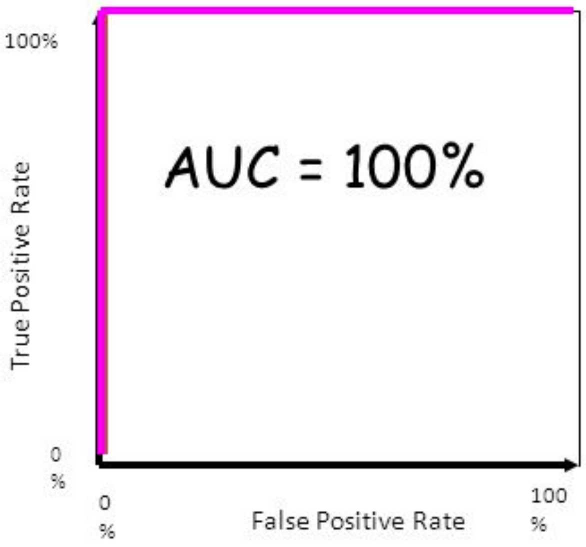


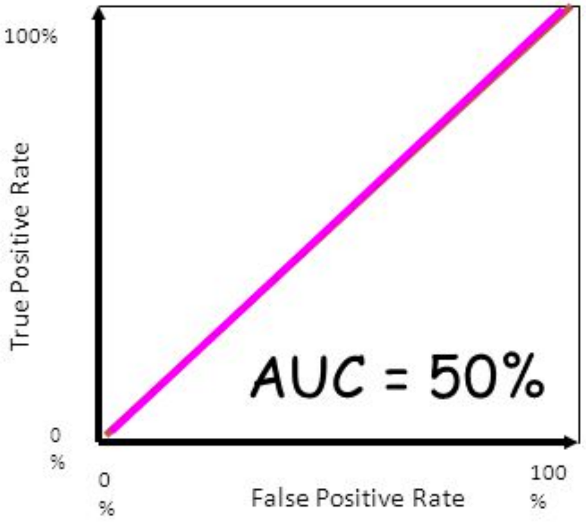
1
best operating point when FN costs 10× FP

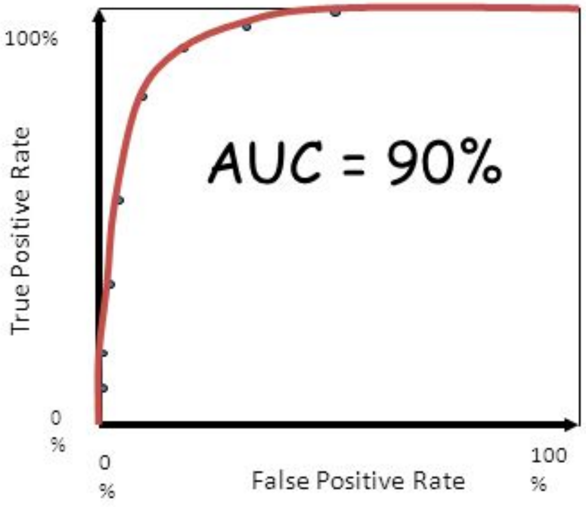
2
best operating point when cost of misclassifying positives and negatives is equal

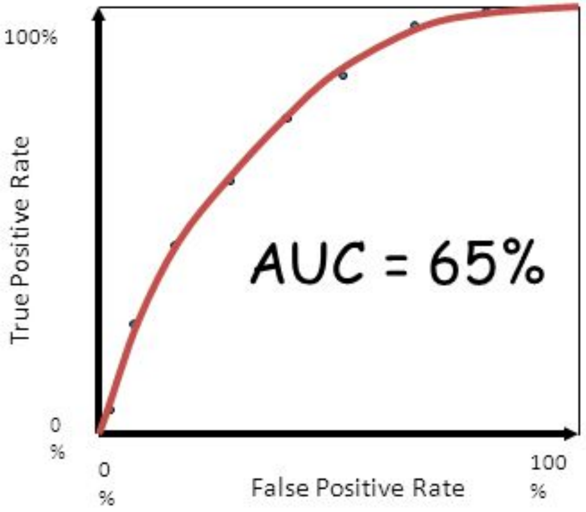
3
best operating point when FP costs 10× FN

AUC for ROC Curve









i.e. don't use with money gain + / loss −

AUC

It doesn't care about absolute values, it only cares about ranking

Get perfect score when:
the predicted score of positive class is higher than negative class in all examples

Metrics for regression task

- Root Mean Square Error (RMSE)
  - Most widely used
  - Emphasize bigger deviations

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

- Mean Absolute Error (MAE)
  - Easiest to interpret

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$