evaluation metrics

- confidence intervals

- For large samples, a good approximation is to assume error is normally distributed

$$\text{CI for Error} = \mathrm{E} \pm Z_\alpha \sqrt{\frac{E\,(1-E)}{N}}$$

- E = error rate
- N = sample size

CI Width	80%	90%	95%	99%
Z_α	1.28	1.64	1.96	2.58

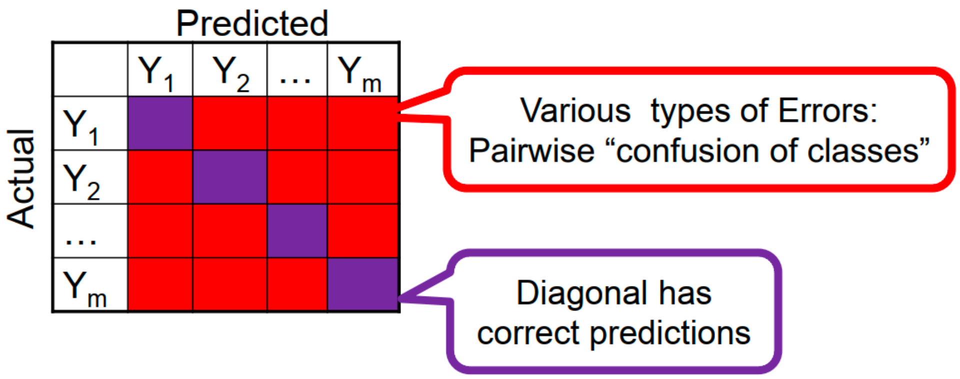
- confusion matrix

for binary classification

	Predicted True	Predicted False
Actually True	True Positive (TP)	False Negative (FN)
Actually False	False Positive (FP)	True Negative (TN)

Number of times algorithm confuses true with false

for multiclass



- class skew

portion of positive samples of all

ROC curves: binary classification

- aspects being plotted:

x-axis: FPR

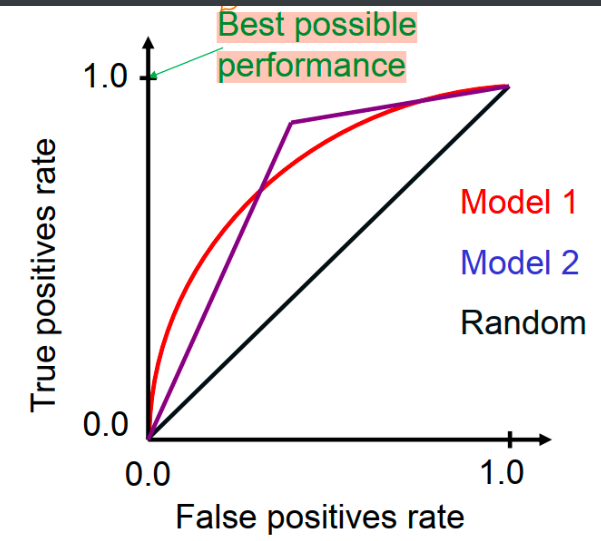
y-axis: TPR

note:

"... rate"'s denominator is always the number of positive / negatives samples!

- best possible performance:

  true positive = 1, false positive = 0, predictions all correct

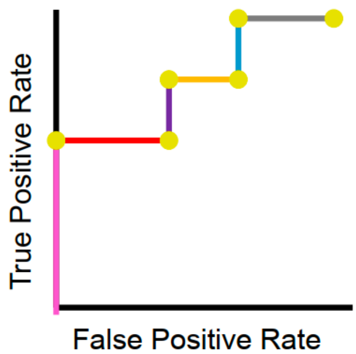


- creating ROC curves

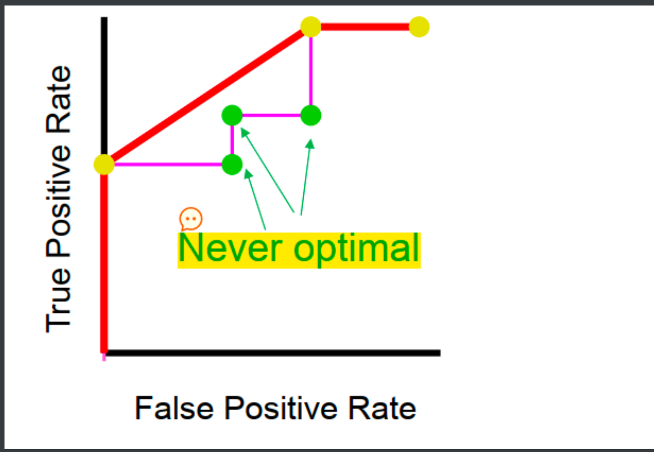
- Approach
  - Sort output values for test set
  - Locate ==boundaries== between examples with different classifications
  - Compute the FPR and TPR for each boundary
  - Plot each (FPR, TPR) point
  - Connect the points

"boundaries" means threshold

Predicted Score	True Label	
0.95	+	
0.93	+	
0.73	+	(0.0, 0.6)
0.63	-	
0.61	-	(0.4, 0.6)
0.52	+	(0.4, 0.8)
0.37	-	(0.6, 0.8)
0.33	+	(0.6, 1.0)
0.12	-	
0.05	-	(1.0, 1.0)

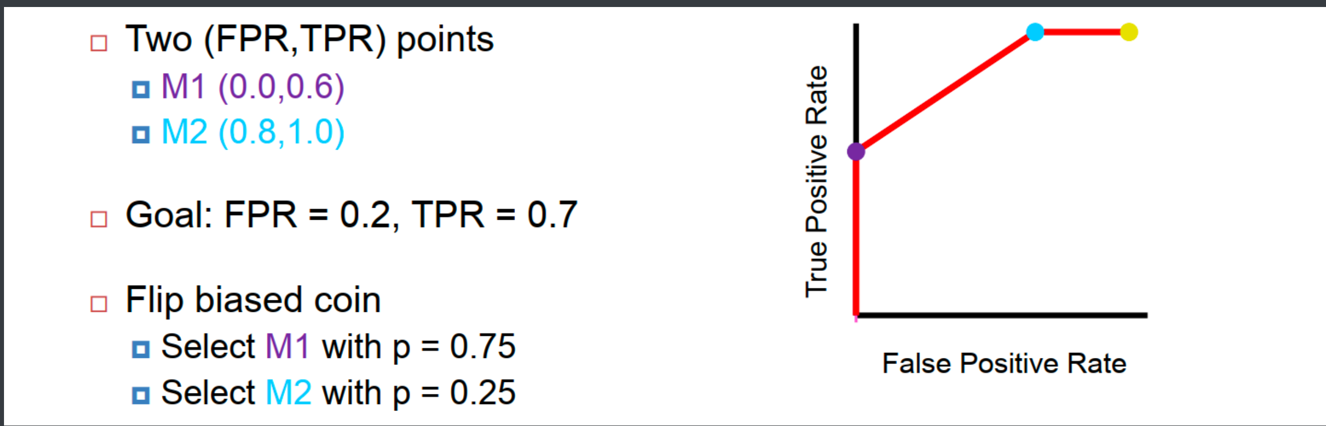


- set of best classifiers: **convex hull**

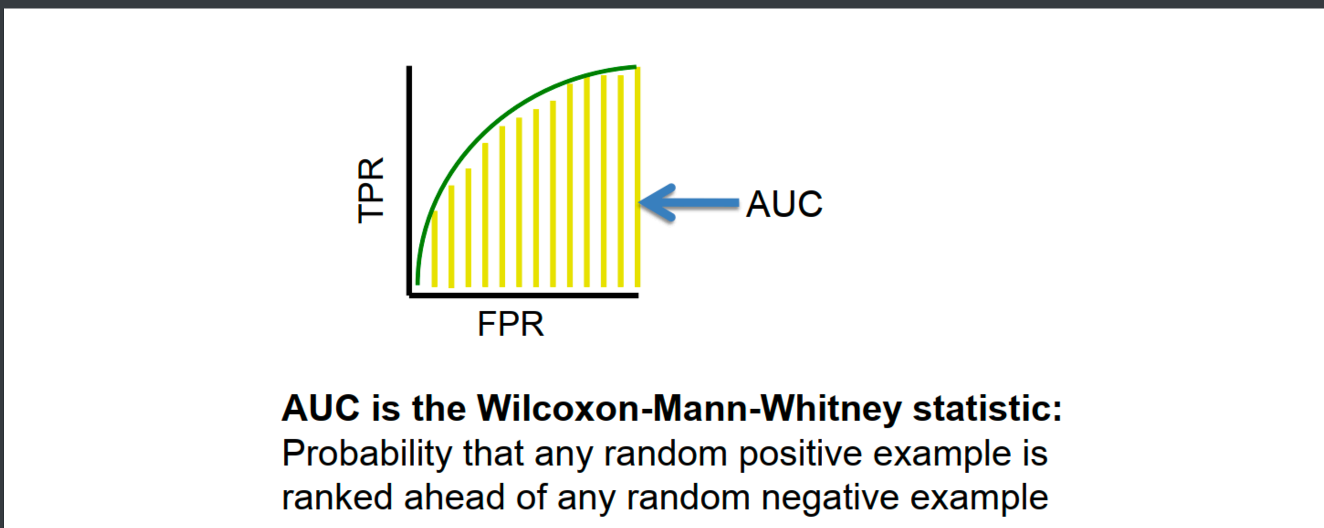


"never optimal" because:

there's always <u>interpolation</u> methods to create a model on the convex hull



- AUC: area under the ROC curve:



- problem: **depends on skew**

suppose we want to **focus on detecting positives**:

if there's lots and lots of negative examples, low FPR could also mean lots of FPs

e.g. not good in cancer diagnosis

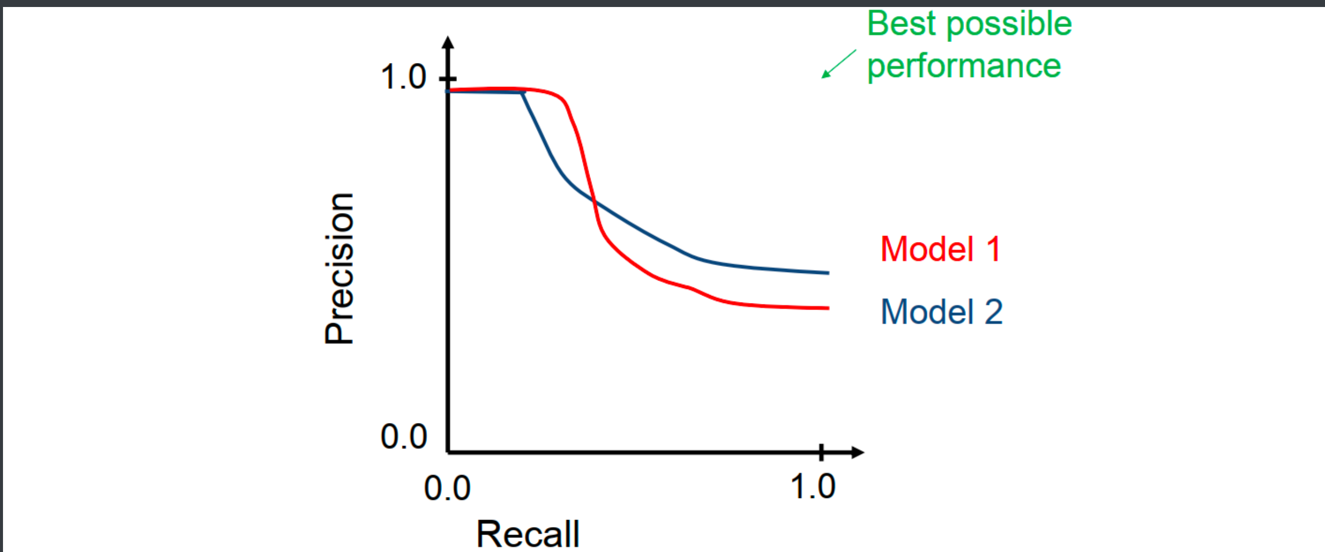
PR curves

- aspects being plotted:

  x-axis: recall = TPR

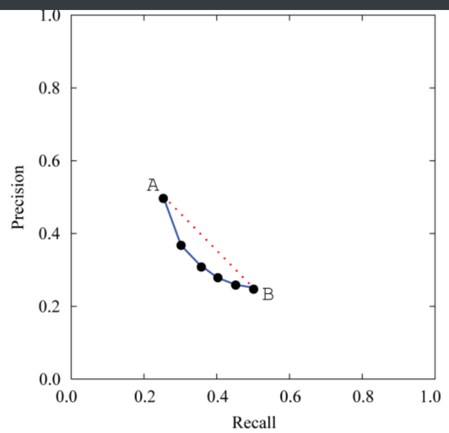
  y-axis: precision = TP / (TP + FP)

- best possible performance:



- !! precision interpolation is counterintuitive !!

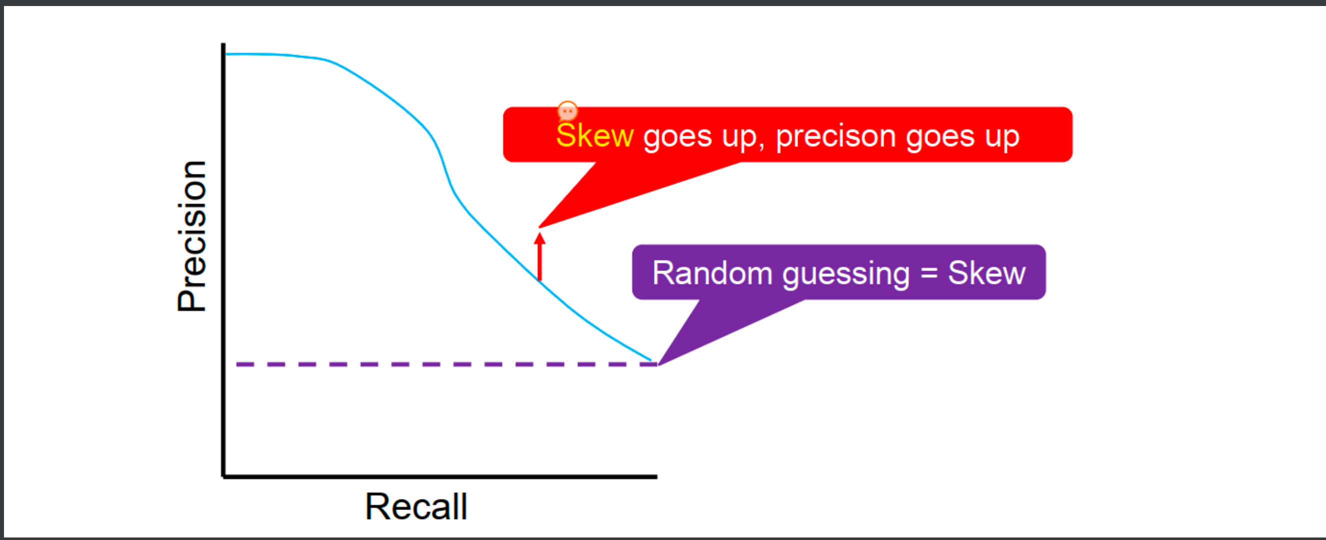
	TP	*FP*	*REC*	*PREC*
A	5	5	0.25	0.5
.	6	10	0.30	0.375
.	7	15	0.35	0.318
.	8	20	0.40	0.286
.	9	25	0.45	0.265
B	10	30	0.5	0.25



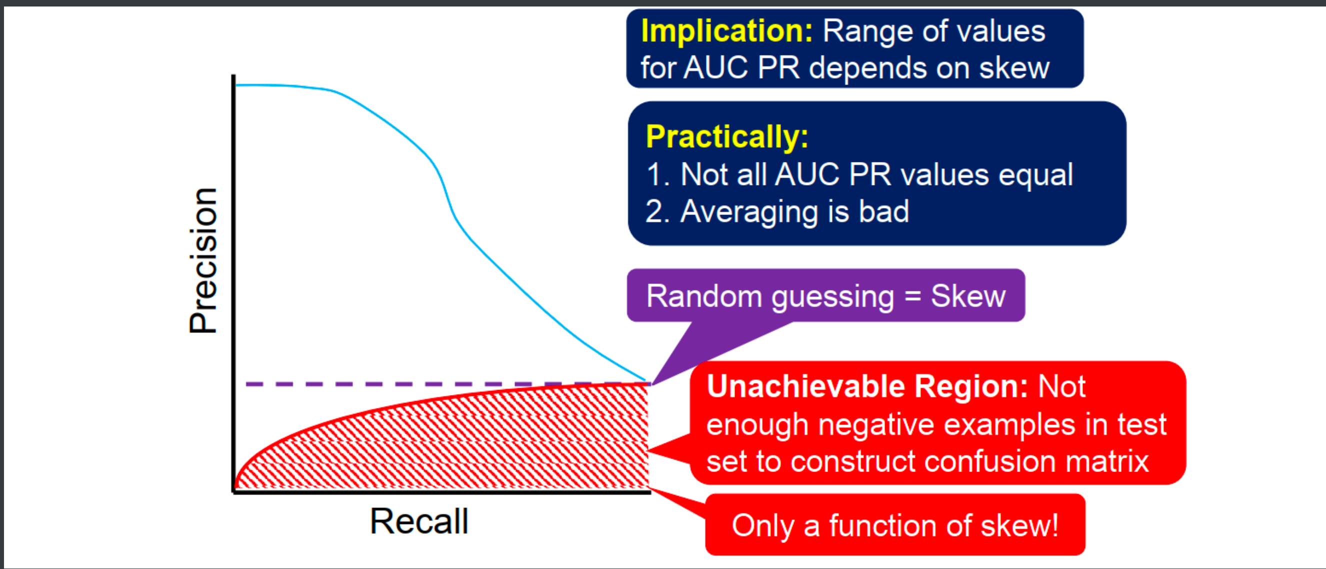
A dataset with 20 positive and 2000 negative examples

  because precision's denominator does not only contain objective values

- correlation between precision & skew



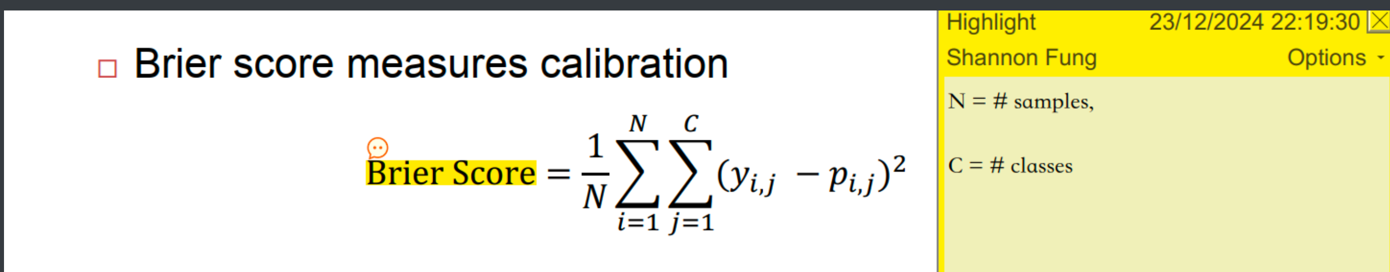
- unachievable region



for probability estimation: calibration

motivation: make predicted probability close to reality

measurement:



calibration method: for example, post process predicted scores

evaluating real-valued outputs

- root-mean square error
- mean absolute error
- relative error

- Relative error = $\frac{\sum_{i=1}^{N}(f(x_i) - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2}$
  - *1 = no better than "always predict mean";*
  - *0 = perfect prediction*