# Confusion Matrix Notes

Correct Classifications: T

Incorrect Classifications: F

| | Predicted Positives $\hat{P}$ | Predicted Negatives $\hat{N}$ |
|---|---|---|
| Actual Positives $P$ | TP<br>$\frac{TP}{P}$: recall or sensitivity ("TPR") or power<br>$\frac{TP}{\hat{P}}$: precision | FN<br>$\frac{FN}{P}$: miss rate (1−sensitivity) ("FNR") |
| Actual Negatives $N$ | FP<br>$\frac{FP}{N}$: fall-out (1−specificity) ("FPR") | TN<br>$\frac{TN}{N}$: specificity ("TNR") |

Summary Metrics:

$\frac{T}{All}$: accuracy

avg(sensitivity, specificity): balanced accuracy

$\left(\frac{n}{\frac{1}{x_1}+\dots\frac{1}{x_n}}\right)$

harmonic_avg(precision, recall): $F_1$

↑ heavily punish any low score, and normalize precision & recall since they have different denominators.

$F_1 = \frac{2}{1/recall + 1/precision} = \frac{2TP}{P+\hat{P}}$

Use sensitivity, specificity, and balanced accuracy for unbalanced data where performance for positives AND negatives is equally important.

$\frac{TP}{P}$: (sensitivity) How good are we at catching all the positive cases there were?

$\frac{TN}{N}$: (specificity) How good are we at catching all the negative cases there were?

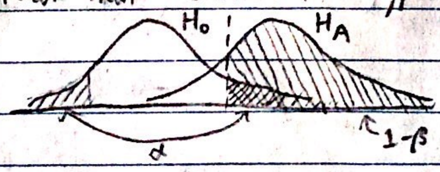
Use recall, precision, and $F_1$ for unbalanced data where performance for positives matters most (find needle in a haystack), like medical data — safety AND efficiency with resources both important.

$\frac{TP}{P}$: (recall) How safe are we at catching all the positive cases there were?

$\frac{TP}{\hat{P}}$: (precision) How efficient is our positive predictor? (consider other side, $\frac{FP}{\hat{P}}$, how wasteful is it?)

$\frac{TP}{P}$: (power): How likely are we to correctly reject the null hypothesis, (so given that the alternative hypothesis is correct) using our decision threshold/classifier?

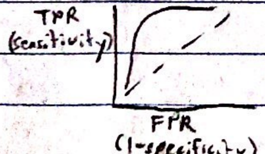
accuracy is fine and intuitive for balanced data.




▨ : $\alpha$ (chance of Type I error; $H_0$ true, thought false)

▧ : $1-\beta$ (power; chance of $H_A$ true, correctly found true)

ROC curve: sensitivity & (1−specificity)
(receiver-operating characteristic)

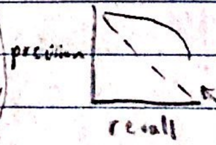



Useful for balanced.

High AUC good.

(despite sensitivity/specificity both being useful for unbalanced with equally important positive/negative, FPR gets inconsistent for unbalanced.)

Precision-Recall curve:




Better than ROC for unbalanced (even when positive vs. negative are both important)

High AUC good.

Probably very noisy.

often doesn't reach 0 precision for balanced data.