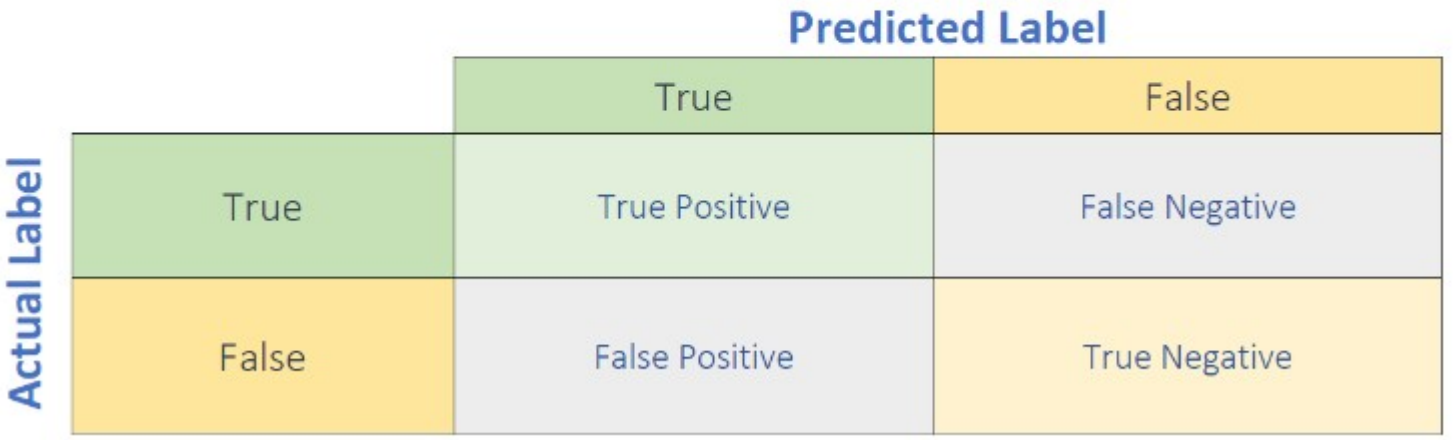
Evaluation Metrics in Machine Learning

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

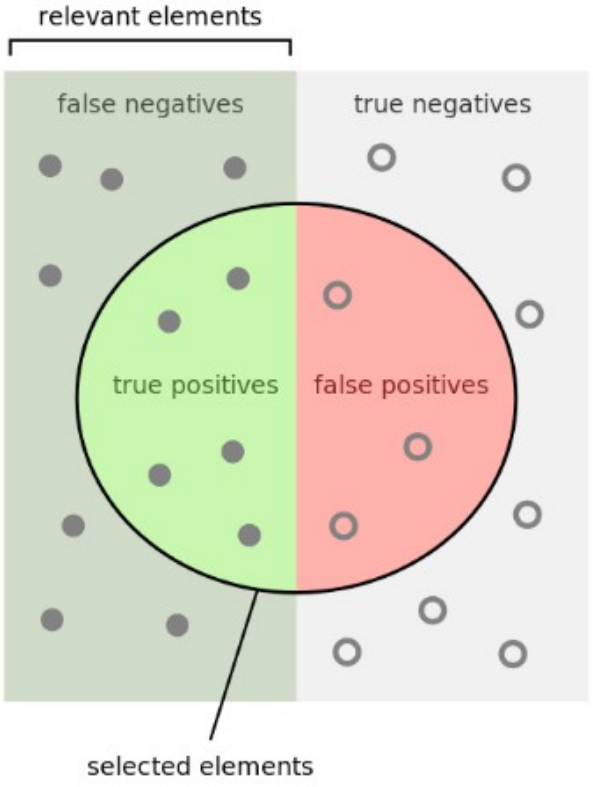
- True Positive (TP) – Actual label is positive and prediction is also positive
- True Negative (TN) – Actual label is negative and prediction is also negative
- False Positive (FP) – Actual label is negative but prediction is positive
- False Negative (FN) – Actual label is positive but prediction is negative



$$Accuracy = \frac{Correct\ Predictions}{All\ Predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{Correct\ Positive\ Predictions}{All\ Positives} = \frac{TP}{TP + FN}$$

$$Precision = \frac{Correct\ Positive\ Predictions}{All\ Positive\ Prediction} = \frac{TP}{TP + FP}$$

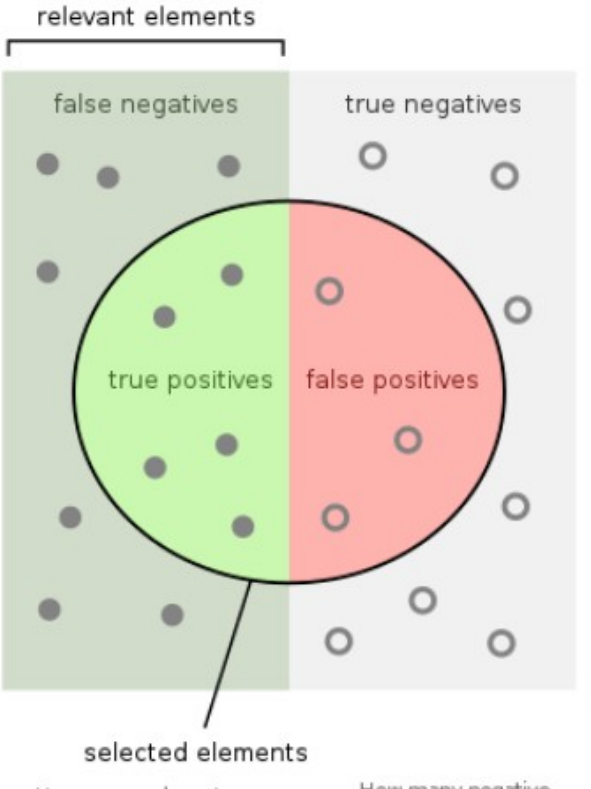


How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$



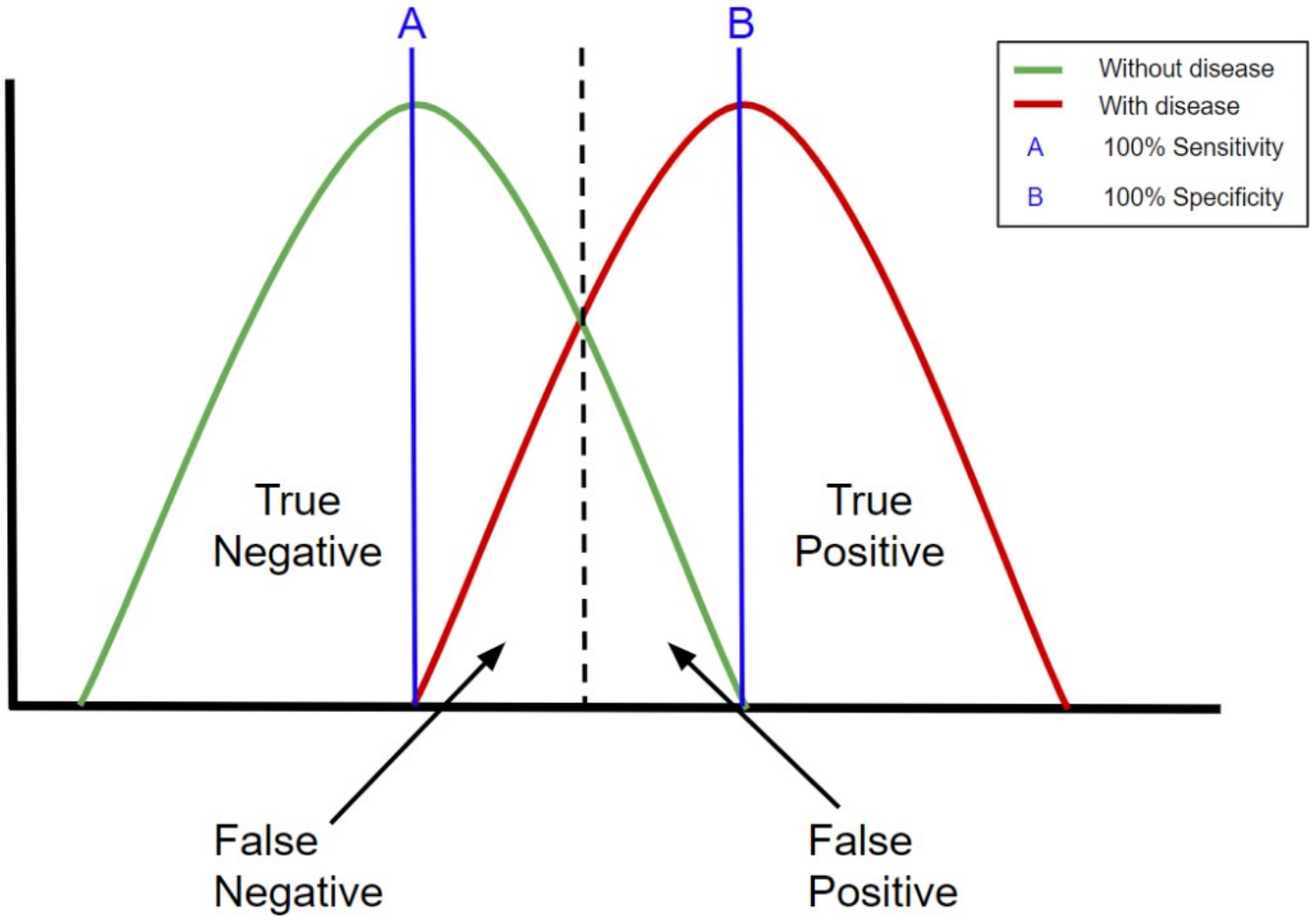
How many relevant items are selected? e.g. How many sick people are correctly identified as having the condition.

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

How many negative selected elements are truly negative? e.g. How many healthy peple are identified as not having the condition.

$$\text{Specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

Sensitivity vs. Specificity



The traditional F-measure or balanced F-score (**F_1 score**) is the harmonic mean of precision and recall:

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{tp}}{\mathrm{tp} + \frac{1}{2}(\mathrm{fp} + \mathrm{fn})}.$$

A more general F score, F_β, that uses a positive real factor β, where β is chosen such that recall is considered β times as important as precision, is:

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}.$$

Confusion matrix [edit]

Consider a group with **P** positive instances and **N** negative instances of some condition. The four outcomes can be formulated in a 2×2 *contingency table* or *confusion matrix*, as well as derivations of several metrics using the four outcomes, as follows:

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Worked example [edit]

A worked example

view · talk · edit

A diagnostic test with sensitivity 67% and specificity 91% is applied to 2030 people to look for a disorder with a population prevalence of 1.48%

		Patients with bowel cancer (as confirmed on endoscopy)				
	Total population (pop.) = 2030	Condition positive	Condition negative	Prevalence = (TP + FN) / pop. = (20 + 10) / 2030 ≈ **1.48%**	Accuracy (ACC) = (TP + TN) / pop. = (20 + 1820) / 2030 ≈ **90.64%**	
Fecal occult blood screen test outcome	Test outcome positive	**True positive** (TP) = 20 (2030 × 1.48% × 67%)	**False positive** (FP) = 180 (2030 × (100% − 1.48%) × (100% − 91%))	Positive predictive value (PPV), precision = TP / (TP + FP) = 20 / (20 + 180) = **10%**	False discovery rate (FDR) = FP / (TP + FP) = 180 / (20 + 180) = **90.0%**	
	Test outcome negative	**False negative** (FN) = 10 (2030 × 1.48% × (100% − 67%))	**True negative** (TN) = 1820 (2030 × (100% − 1.48%) × 91%)	False omission rate (FOR) = FN / (FN + TN) = 10 / (10 + 1820) ≈ **0.55%**	Negative predictive value (NPV) = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.45%**	
		True positive rate (TPR), recall, **sensitivity** = TP / (TP + FN) = 20 / (20 + 10) ≈ **66.7%**	False positive rate (FPR), fall-out, probability of false alarm = FP / (FP + TN) = 180 / (180 + 1820) = **9.0%**	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$ = (20 / 30) / (180 / 2000) ≈ **7.41**	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$ ≈ **20.2**	F_1 score = 2 × $\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$ ≈ **0.174**
		False negative rate (FNR), miss rate = FN / (FN + TP) = 10 / (20 + 10) ≈ **33.3%**	**Specificity**, selectivity, true negative rate (TNR) = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$ = (10 / 30) / (1820 / 2000) ≈ **0.366**		