UNC CHARLOTTE

The WILLIAM STATES LEE COLLEGE of ENGINEERING

Introduction to ML
Lecture 8: Classifier Evaluation

Hamed Tabkhi

Department of Electrical and Computer Engineering,

University of North Carolina Charlotte (UNCC)

htabkhiv@uncc.edu

UNC CHARLOTTE

Overfitting

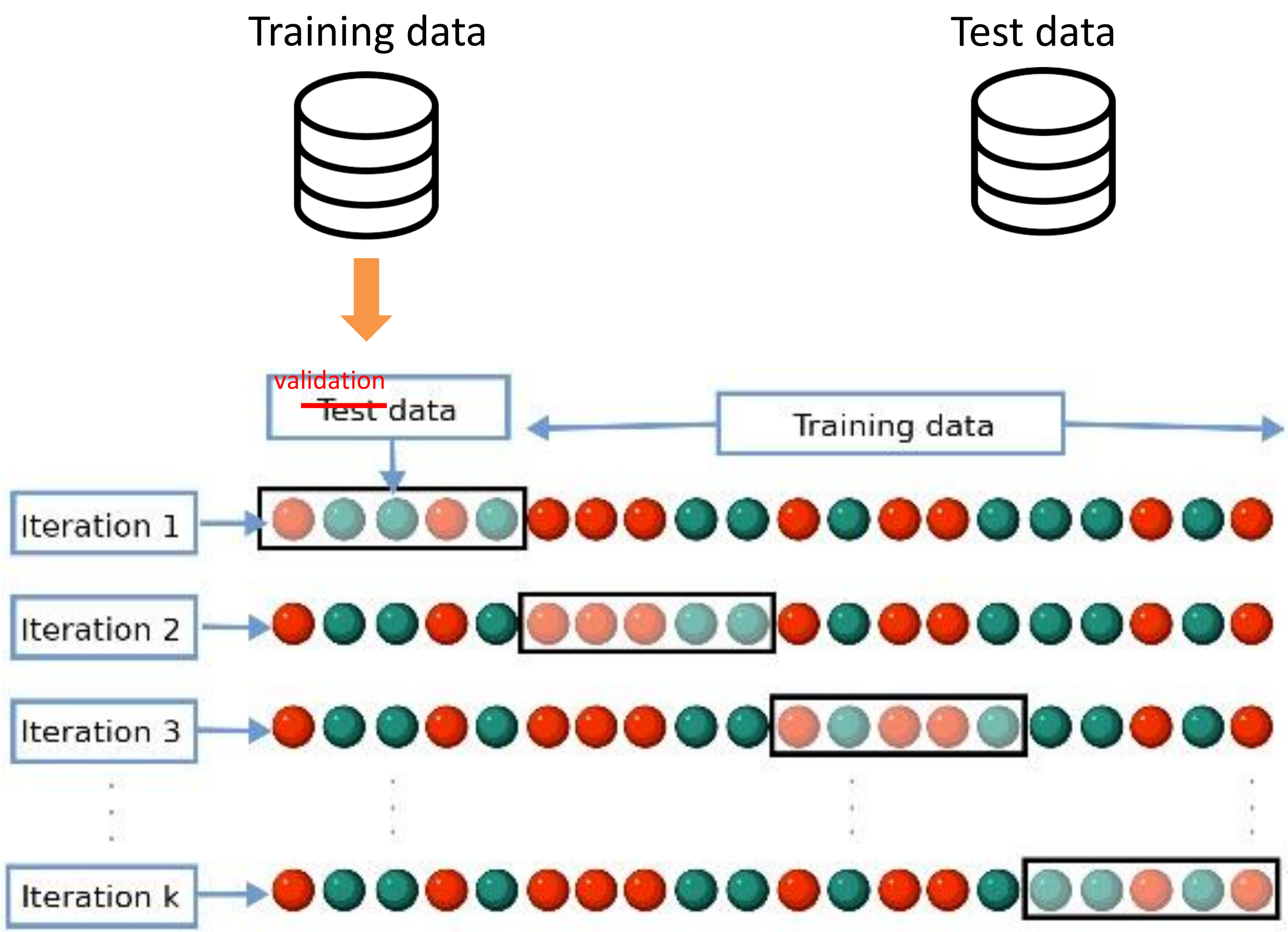
- Overfitting:
- A classifier that performs well on the training examples, but poorly on new examples.
- Training and testing on the same data will generally produce a good classifier (for this dataset) with high overfitting. (Never do this!)

- To avoid overfitting:
- Use cross-validation

UNC CHARLOTTE

K-fold Cross-validation

- Use cross-validation to avoid overfitting



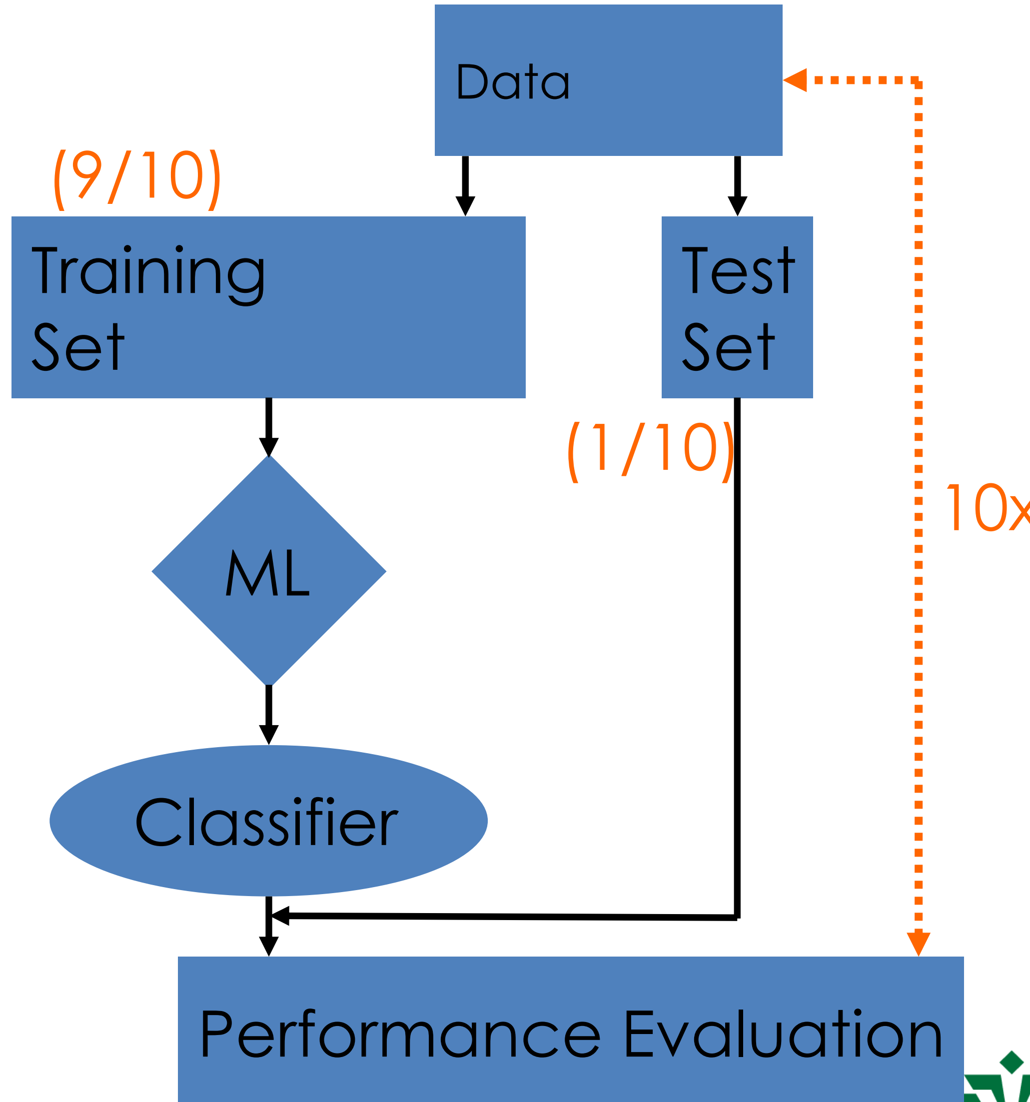
K-fold Cross-validation

1. Split the entire data randomly into K folds (value of K shouldn't be too small or too high, ideally we choose 5 to 10 depending on the data size). The higher value of K leads to less biased model (but large variance might lead to over-fit), where as the lower value of K is similar to the train-test split approach we saw before.
2. Then fit the model using the K-1 (K minus 1) folds and validate the model using the remaining Kth fold. Note down the scores/errors.
3. Repeat this process until at least every K-fold serve as the test set.
4. For the final accuracy measurement the average of your recorded scores. That will be the performance metric for the model.

The most common strategies

- Cross-Validation (e.g., 10 fold)

Training data (all data)



UNC CHARLOTTE

Evaluation metrics

- Accuracy, error rate
  - Accuracy is the percent of correct classifications
    - **Accuracy** = Correct Predictions / Total Predictions
  - Error rate is the percent of incorrect classifications
  - Accuracy = 1 – Error rate
- Problems with the accuracy
  - Assumes equal costs for misclassification
  - Assumes relatively uniform class distribution
    - E.g. imbalanced dataset. Consider 95 negative samples and 5 positive samples. Classifying all samples as negative in this case gives 0.95 accuracy score.

UNC CHARLOTTE

Evaluation metrics

	Predicted Y	Predicted N
Actually Y	True Positive	False Negative
Actually N	False Positive	True Negative

Evaluation metrics

True Positive: we correctly detect the class

False Positive: we predict a target class for a negative sample - cause false alarm

	Predicted Y	Predicted N
Actually Y	True Positive	False Negative
Actually N	False Positive	True Negative

Evaluation metrics

True Positive: we correctly detect the class

False Positive: we predict a target class for a negative sample
- Cause false alarm

False Negative: We were not able to predict a correct class for a positive sample
- Can be very bad in many applications

	Predicted Y	Predicted N
Actually Y	True Positive	False Negative
Actually N	False Positive	True Negative

UNC CHARLOTTE

Evaluation metrics

True Positive: we correctly detect the class

False Positive: we predict a target class for a negative sample
- Cause false alarm

False Negative: We were not able to predict a correct class for a positive sample
- Can be very bad in many applications

True Negative?:

	Predicted Y	Predicted N
Actually Y	True Positive	False Negative
Actually N	False Positive	True Negative

Evaluation metrics

recall, sensitivity, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

How much of the real 'Yes' cases are detected? How well can it detect the condition?

specificity, selectivity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

How much of the real 'No' cases are correctly classified? How well can it rule out the condition?

precision or positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

	Predicted Y	Predicted N
Actually Y	True Positive	False Negative
Actually N	False Positive	True Negative

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

UNC CHARLOTTE

Evaluation metrics

- Previous example: 95 negative samples and 5 positive samples
  - Classifying all samples as negative in this case gives 0.95 accuracy score.

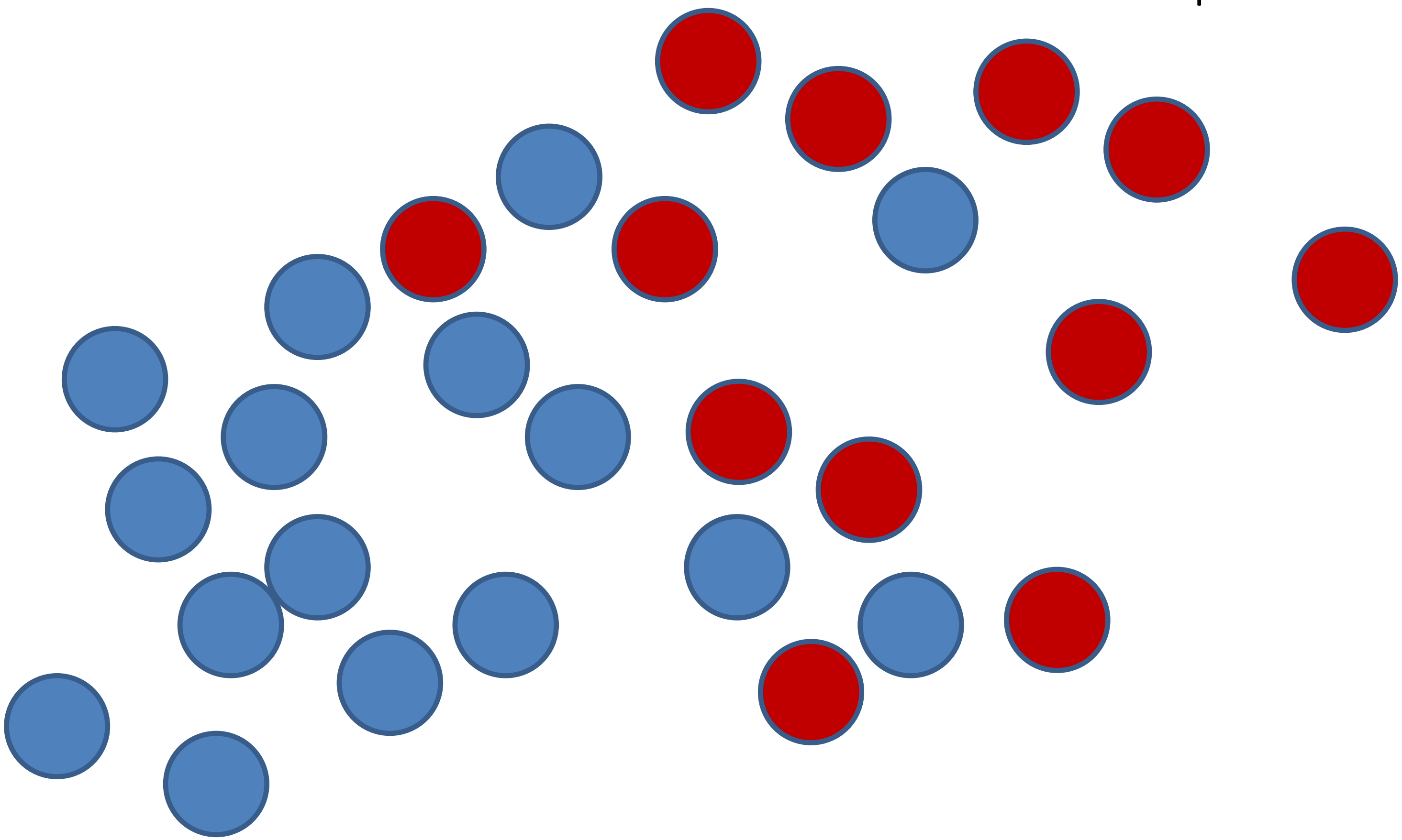
$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

YES

NO

Detect cancer cases

samples



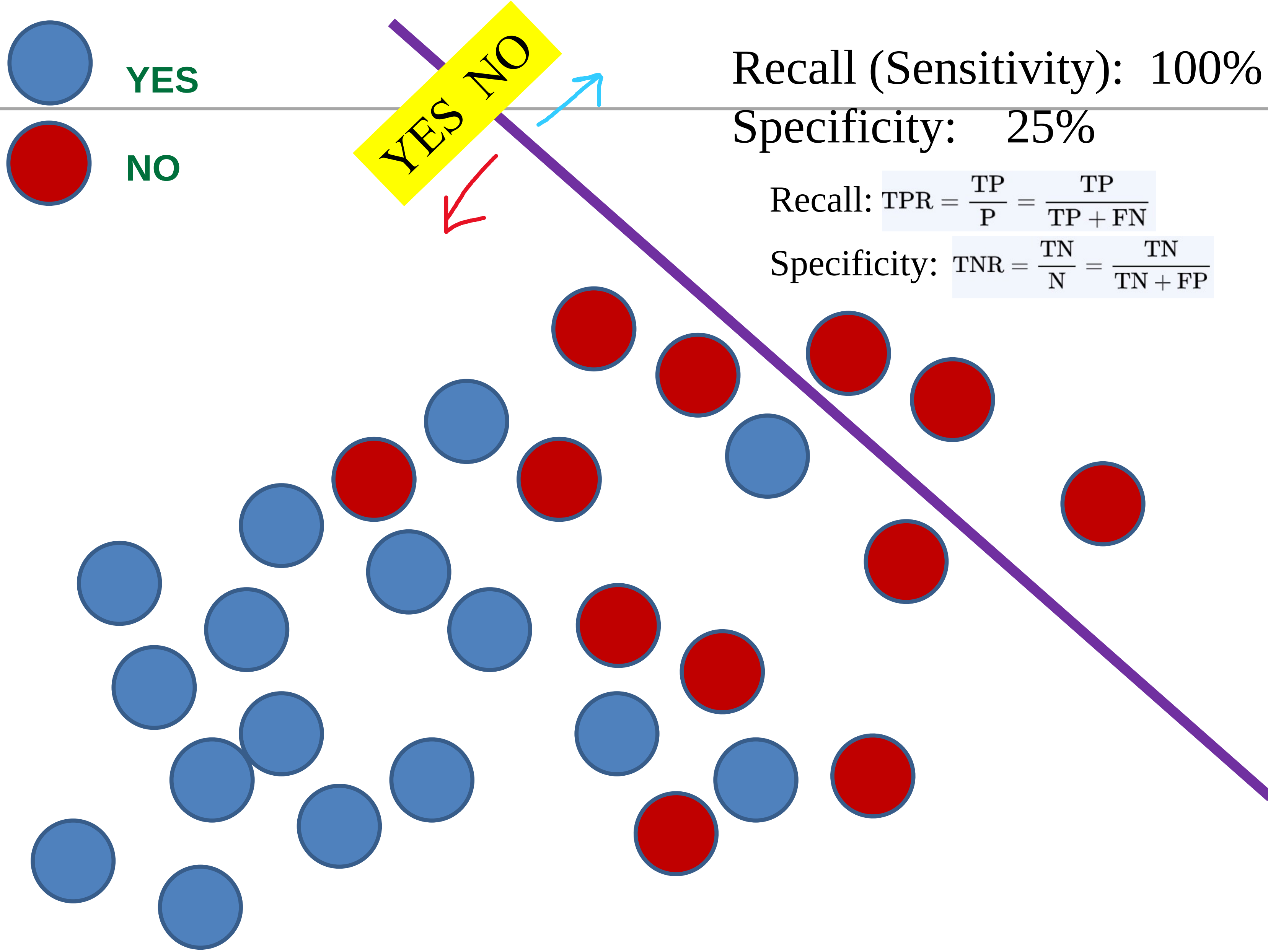
UNC CHARLOTTE

Recall (Sensitivity): 100%

Specificity: 25%

Recall: $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$

Specificity: $\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$



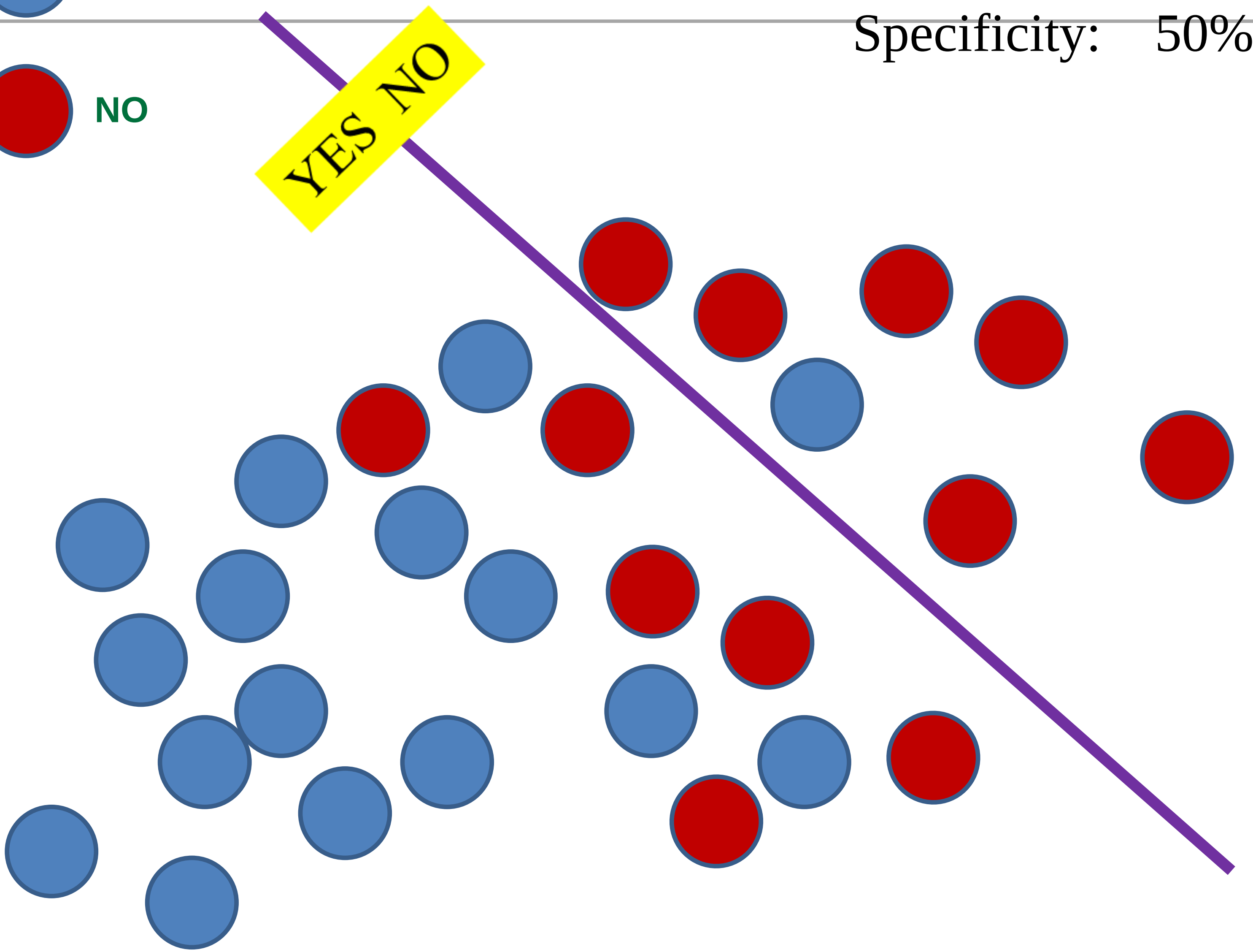
UNC CHARLOTTE

YES

NO

Recall: 93.8%

Specificity: 50%



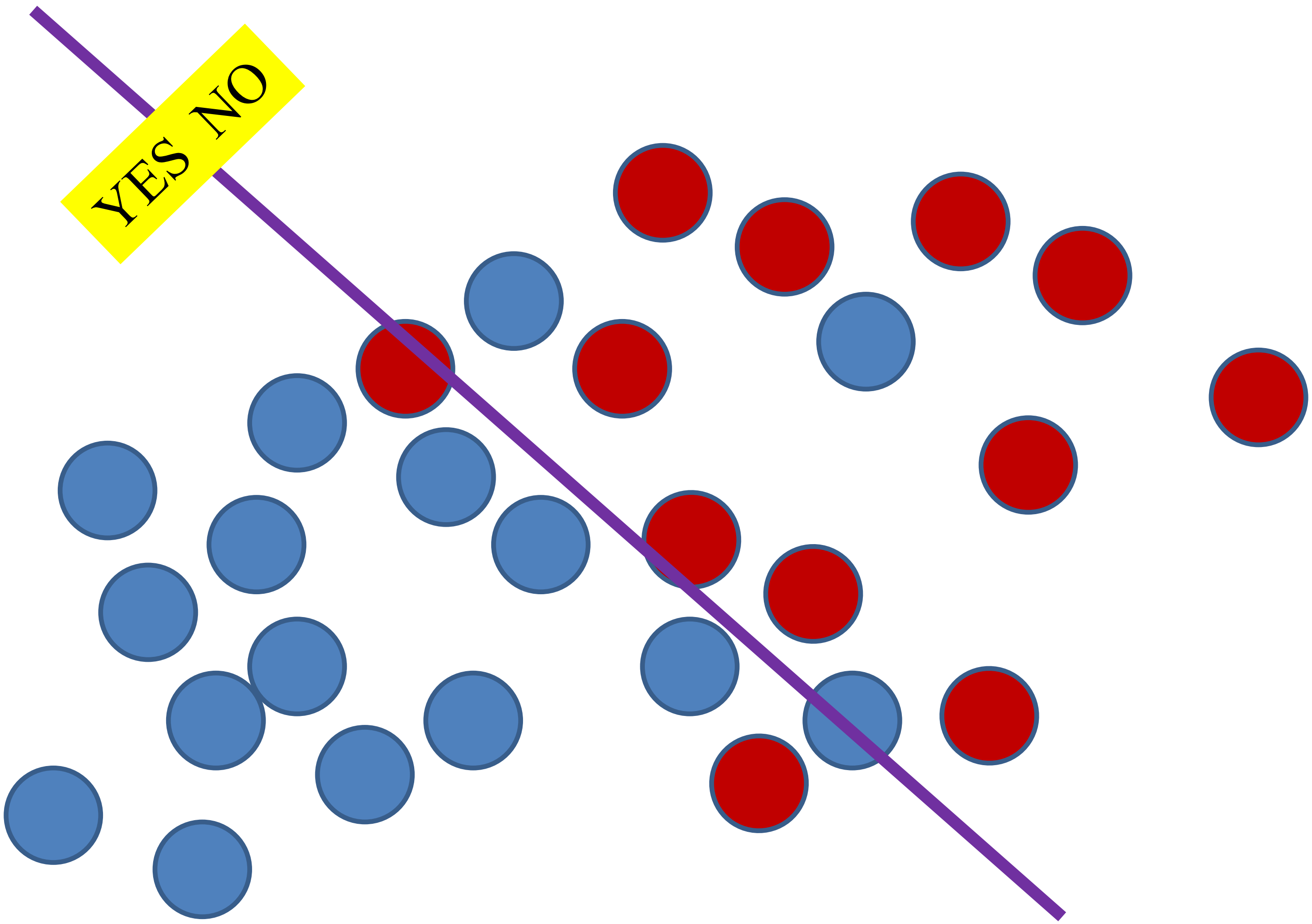
UNC CHARLOTTE

YES

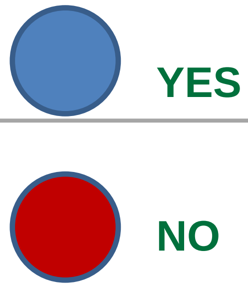
NO

Recall: 81.3%

Specificity: 83.3%

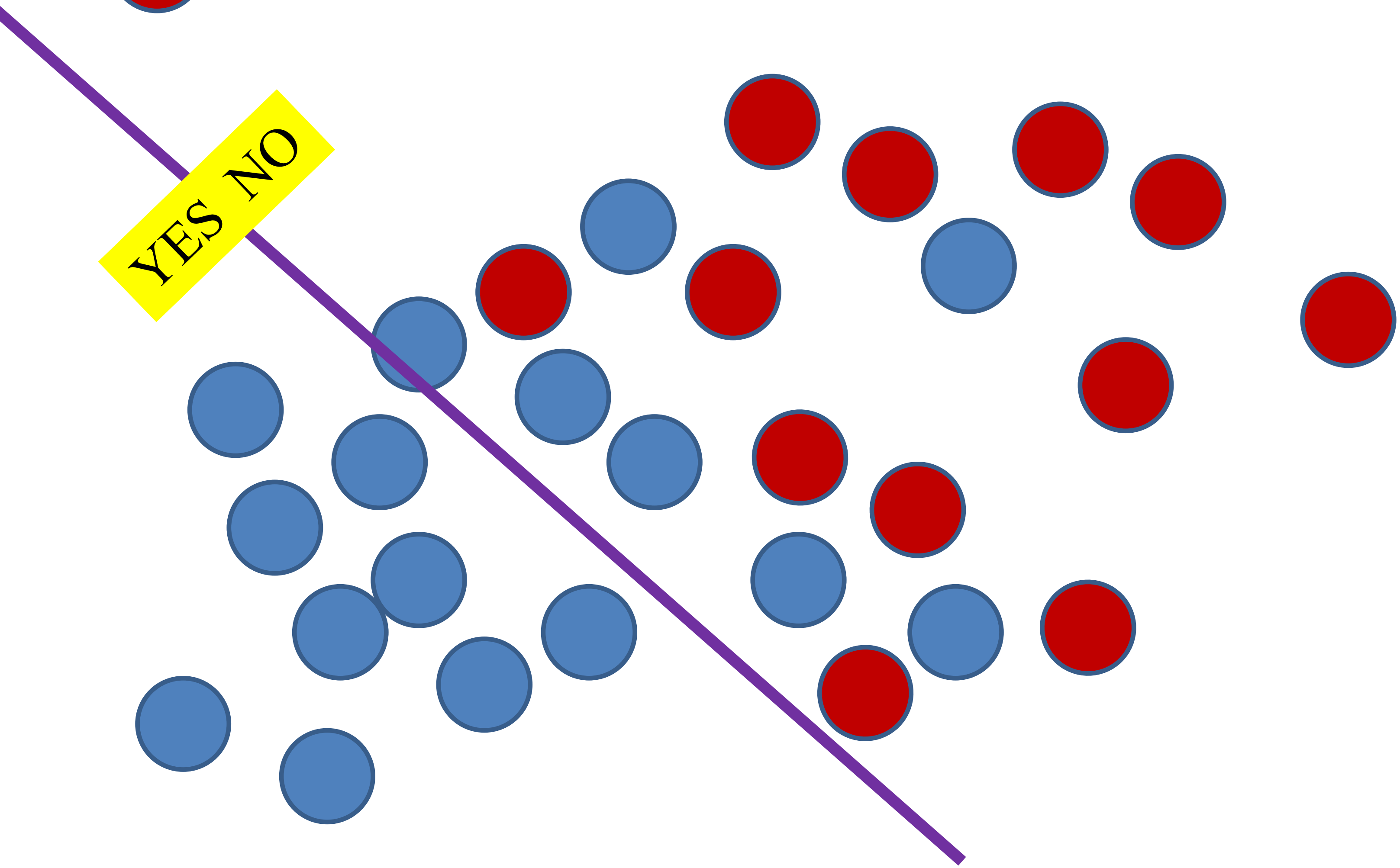


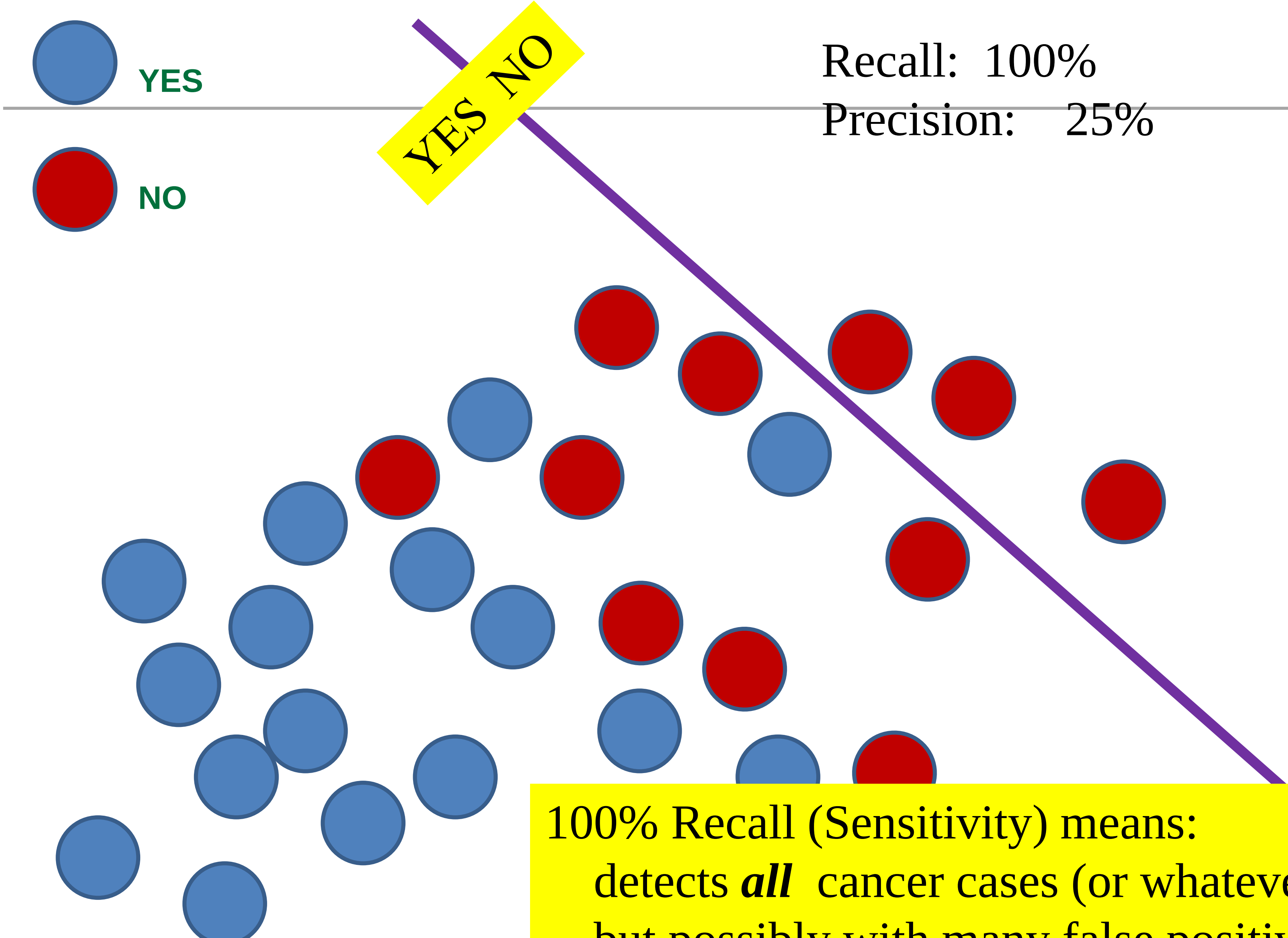
UNC CHARLOTTE



Recall: 56.3%

Precision: 100%

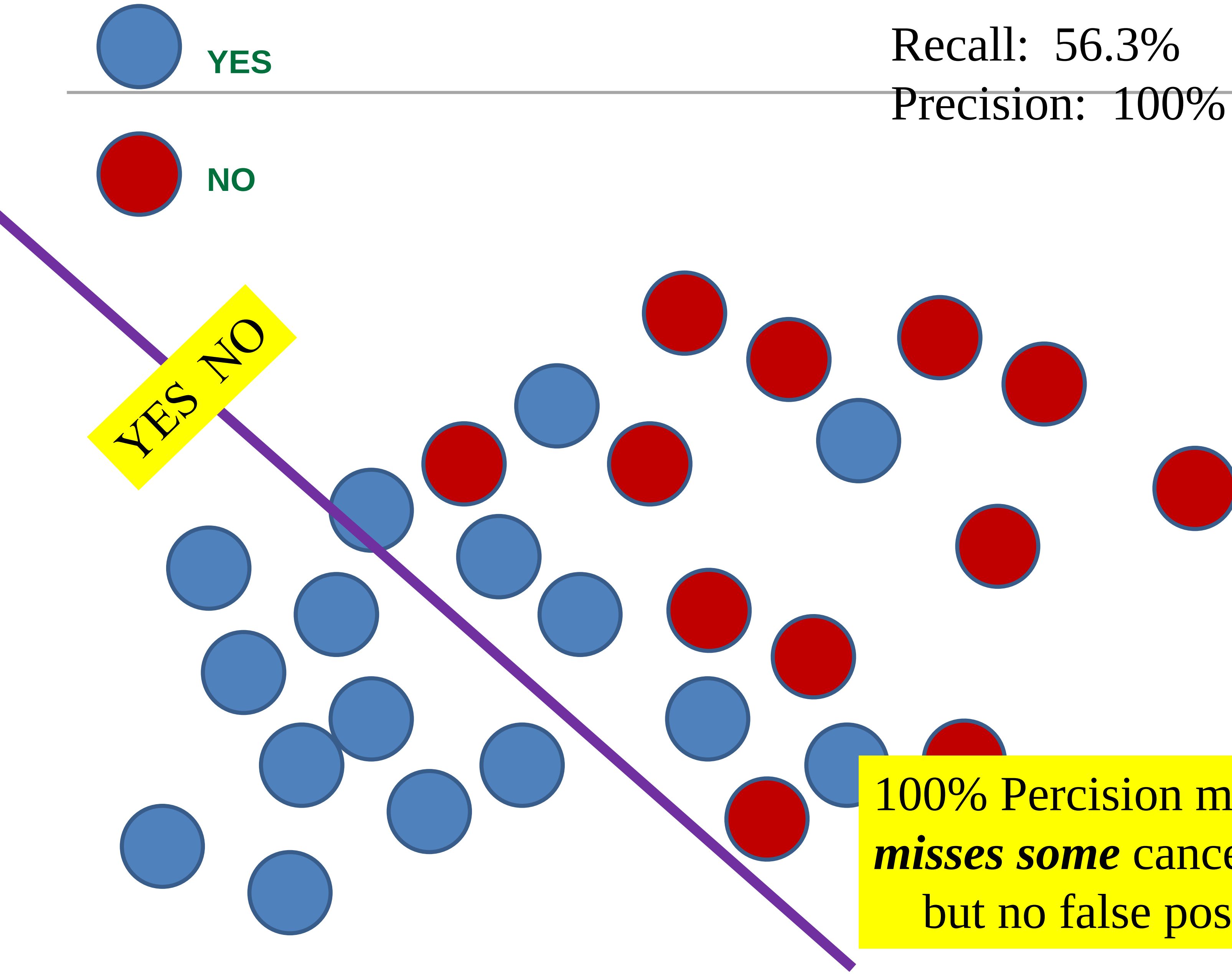




Recall: 100%
Precision: 25%

100% Recall (Sensitivity) means:
detects ***all*** cancer cases (or whatever)
but possibly with many false positives

Recall: 56.3%
Precision: 100%



100% Percision means:
misses some cancer cases
but no false positives

Evaluation metrics

- Confusion matrix (> 2 classes)

Predicted class

sum of a corresponding row

Actual class	1	2	3	4	5	6	7	8	9	Acc
1	**137**	13	3	0	0	1	1	0	0	0.89
2	1	**55**	1	0	0	0	0	6	1	0.86
3	2	4	**84**	0	0	0	1	1	2	0.89
4	3	0	1	**153**	5	2	1	1	1	0.92
5	0	0	3	0	**44**	2	2	1	2	0.82
6	0	0	2	1	4	**35**	0	0	1	0.81
7	0	0	0	0	0	0	**61**	2	2	0.94
8	0	0	0	1	0	0	0	**69**	3	0.95
9	0	0	0	0	0	0	0	2	**26**	0.93
										0.89

What is the total number of test samples of each class?

Evaluation metrics

- Confusion matrix (> 2 classes)

Predicted class

Actual class

	1	2	3	4	5	6	7	8	9	Acc
1	**137**	13	3	0	0	1	1	0	0	0.89
2	1	**55**	1	0	0	0	0	6	1	0.86
3	2	4	**84**	0	0	0	1	1	2	0.89
4	3	0	1	**153**	5	2	1	1	1	0.92
5	0	0	3	0	**44**	2	2	1	2	0.82
6	0	0	2	1	4	**35**	0	0	1	0.81
7	0	0	0	0	0	0	**61**	2	2	0.94
8	0	0	0	1	0	0	0	**69**	3	0.95
9	0	0	0	0	0	0	0	2	**26**	0.93
										0.89

What is the TP for each class?

Each diagonal element corresponds to the TP of a class

UNC CHARLOTTE

Evaluation metrics

- Confusion matrix (> 2 classes)

Predicted class

Actual class	1	2	3	4	5	6	7	8	9	Acc
1	**137**	13	3	0	0	1	1	0	0	0.89
2	1	**55**	1	0	0	0	0	6	1	0.86
3	2	4	**84**	0	0	0	1	1	2	0.89
4	3	0	1	**153**	5	2	1	1	1	0.92
5	0	0	3	0	**44**	2	2	1	2	0.82
6	0	0	2	1	4	**35**	0	0	1	0.81
7	0	0	0	0	0	0	**61**	2	2	0.94
8	0	0	0	1	0	0	0	**69**	3	0.95
9	0	0	0	0	0	0	0	2	**26**	0.93
										0.89

What is the total number of FN for a class?

The sum of values in the corresponding **row** (excluding the TP)

UNC CHARLOTTE

Evaluation metrics

- Confusion matrix (> 2 classes)

Predicted class

Actual class

	1	2	3	4	5	6	7	8	9	Acc
1	**137**	13	3	0	0	1	1	0	0	0.89
2	1	**55**	1	0	0	0	0	6	1	0.86
3	2	4	**84**	0	0	0	1	1	2	0.89
4	3	0	1	**153**	5	2	1	1	1	0.92
5	0	0	3	0	**44**	2	2	1	2	0.82
6	0	0	2	1	4	**35**	0	0	1	0.81
7	0	0	0	0	0	0	**61**	2	2	0.94
8	0	0	0	1	0	0	0	**69**	3	0.95
9	0	0	0	0	0	0	0	2	**26**	0.93
										0.89

What is the total number of FP for a class?

The sum of values in the corresponding **column** (excluding the TP)

Evaluation metrics

Confusion matrix

	PREDICTED					
ACTUAL		A	B	C	D	E
	A	TP_A	E_{AB}	E_{AC}	E_{AD}	E_{AE}
	B	E_{BA}	TP_B	E_{BC}	E_{BD}	E_{BE}
	C	E_{CA}	E_{CB}	TP_C	E_{CD}	E_{CE}
	D	E_{DA}	E_{DB}	E_{DC}	TP_D	E_{DE}
	E	E_{EA}	E_{EB}	E_{EC}	E_{ED}	TP_E

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Recall A? Recall A = Sensitivity A = TP_A / (TP_A + E_{AB} + E_{AC} + E_{AD} + E_{AE})

Recall B? Recall B = Sensitivity B = TP_B / (TP_B + E_{BA} + E_{BC} + E_{BD} + E_{BE})

Model overall performance = average of the class-wise recall

UNC CHARLOTTE

Evaluation metrics

Confusion matrix

		PREDICTED				
ACTUAL		A	B	C	D	E
	A	TP_A	E_{AB}	E_{AC}	E_{AD}	E_{AE}
	B	E_{BA}	TP_B	E_{BC}	E_{BD}	E_{BE}
	C	E_{CA}	E_{CB}	TP_C	E_{CD}	E_{CE}
	D	E_{DA}	E_{DB}	E_{DC}	TP_D	E_{DE}
	E	E_{EA}	E_{EB}	E_{EC}	E_{ED}	TP_E

precision or positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

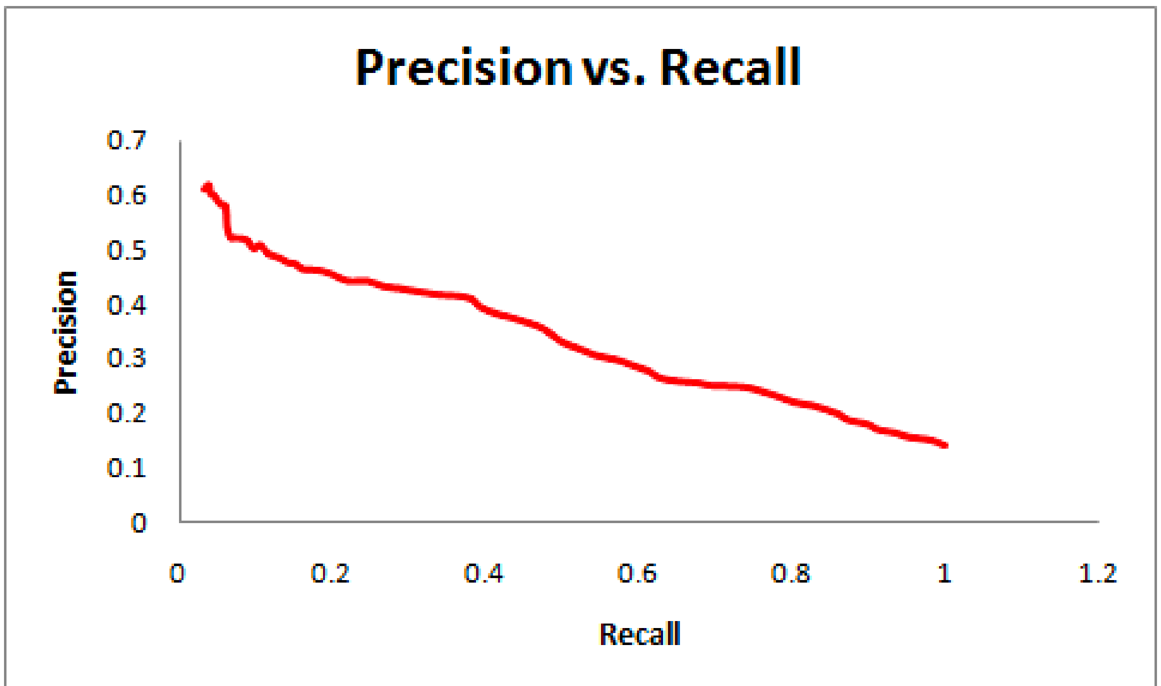
Precision A? $\quad$ $\text{Precision A} = TP_A / (TP_A + E_{BA} + E_{CA} + E_{DA} + E_{EA})$

Precision B? $\quad$ $\text{Precision B} = TP_B / (TP_B + E_{AB} + E_{CB} + E_{DB} + E_{EB})$

Model overall performance = average of the class-wise precision

Evaluation metrics

- Precision vs. Recall
  - In practice, one always needs to make a compromise between these two metrics: by increasing Recall, we decrease (though unwillingly) Precision, and vice versa



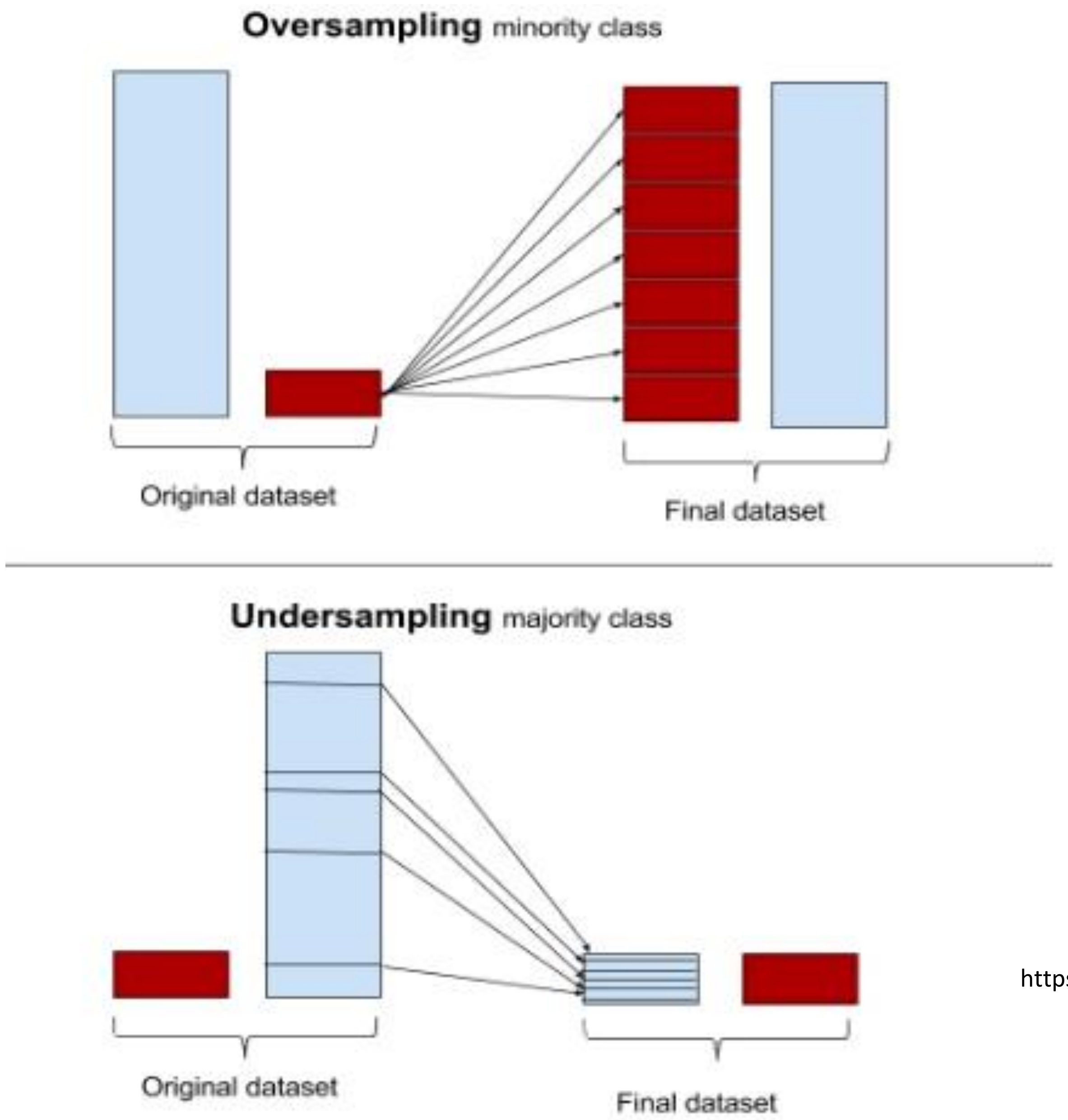
https://groups.csail.mit.edu/cb/struct2net/webserver/images/prec-v-recall-v2.png

Imbalanced data?

- Solutions
  - Oversampling: re-sampling of data from minority class
  - Under-sampling: randomly eliminate samples from majority class
  - Synthesizing new data points for minority class
    - Take averages of samples in minority class
    - Add small noise to samples in minority class

  - We will talk about this more in deep learning

Imbalanced data?



https://www.svds.com/learning-imbalanced-classes/

UNC CHARLOTTE