Performance measures and model comparison

Occam's Razor

- If two models are generally similar in terms of their error statistics and other diagnostics, you should prefer the one that is simpler and/or easier to understand

Regression

Regression

- Simple approach: compare errors (eg RMSE)
- Or R-squared (variance explained). If the models do not have the same complexity, then use adjusted R-squared
- There is no absolute standard for a "good" value of adjusted R-squared

Categorization

		Condition (as determined by "Gold standard")	
	Total population	Condition positive	Condition negative
Test outcome	Test outcome positive	True positive	False positive (Type I error)
	Test outcome negative	False negative (Type II error)	True negative

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

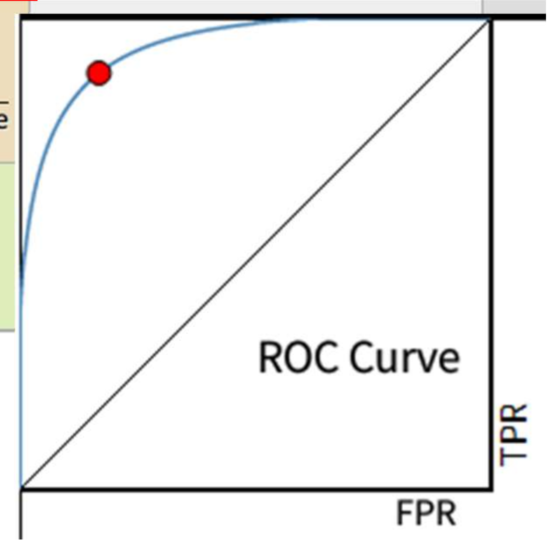
Why do we need different performance measures

Example with lots of bias

A simple model says ‘A’ all the time

Accuracy: 90%

		Condition (as determined by "Gold standard")		
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$
	Test outcome negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	True positive rate (TPR), Sensitivity, Recall = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	

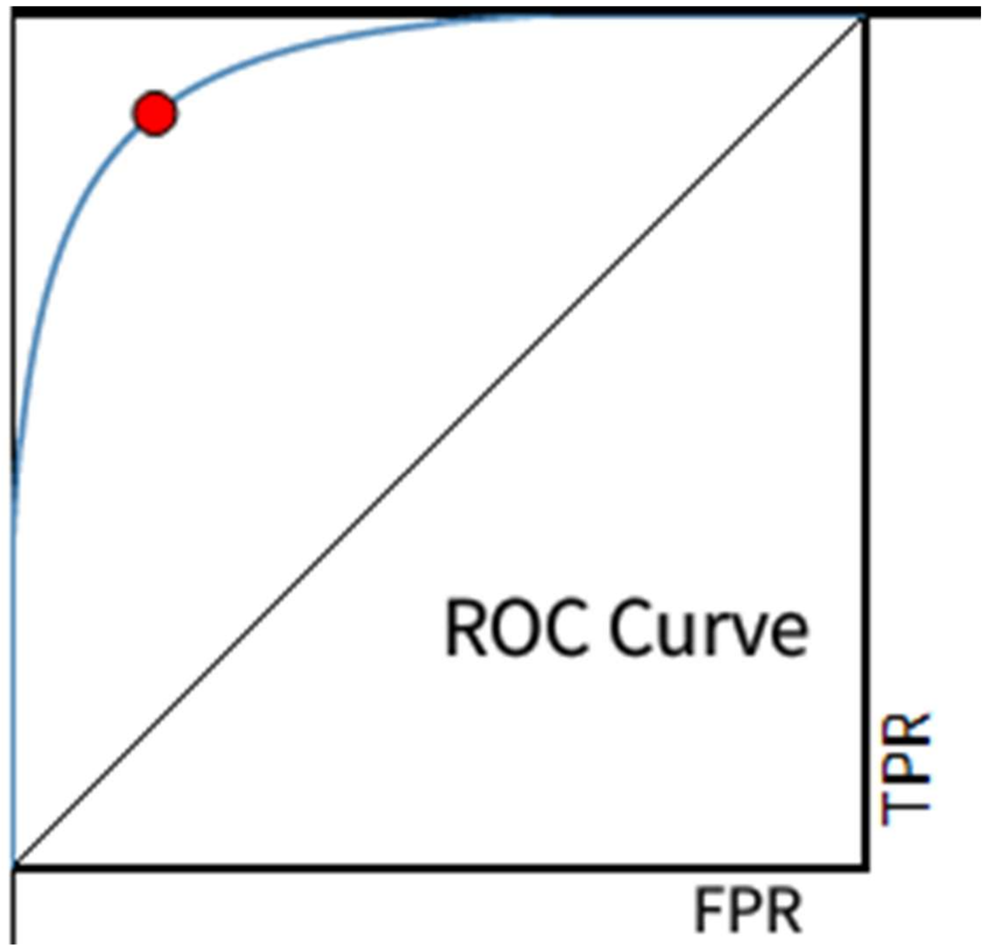


ROC curve

y-axis is true positive rate, and the *x*-axis is false positive rate

Interpretation:

Pick a random negative and a random positive example; The AUC gives you the probability that your classifier assigns a higher score to the positive example (ie, ranks the positive higher than the negative).



ROC curve

- The most common method for combining sensitivity and specificity into a single value uses the receiver operating characteristic (ROC) curve.

- The ROC curve is useful for determining alternate cutoffs for class probabilities

Precision and recall

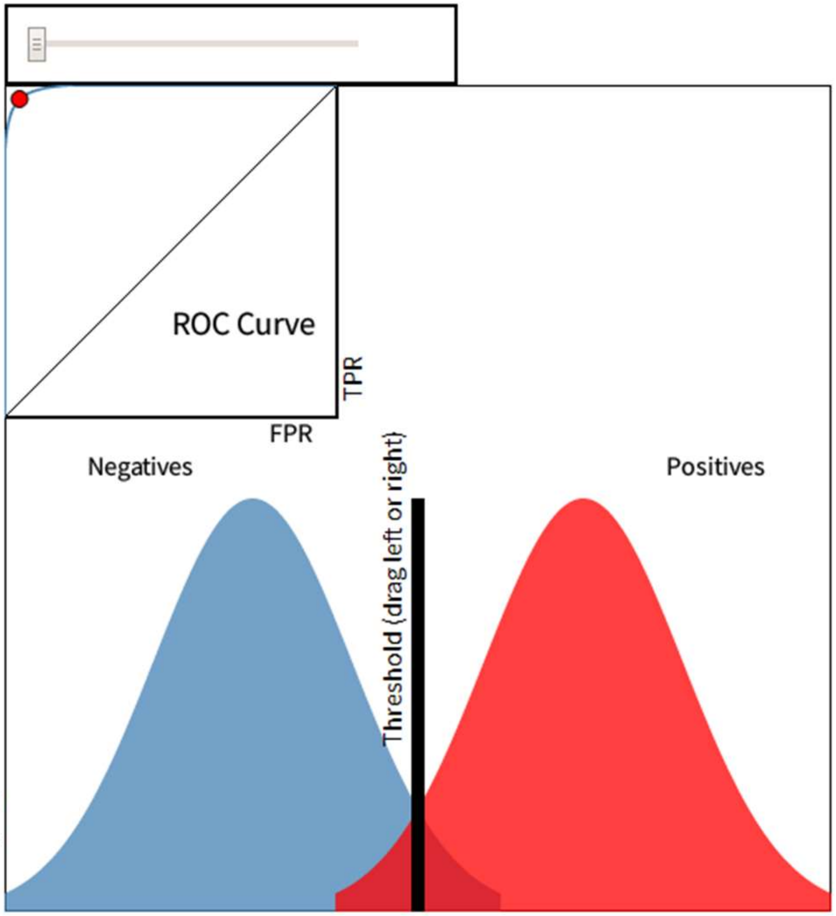
- Origins: information retrieval
- Precision is the probability that a (randomly selected) retrieved document is relevant.

- Recall is the probability that a (randomly selected) relevant document is retrieved in a search.
- They are balanced: you can increase one at the cost of the other

		Condition (as determined by "Gold standard")			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	=
	Test outcome negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Ne =
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	True positive rate (TPR), Sensitivity, Recall $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	D
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$	

		Condition (as determined by "Gold standard")		
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$
	Test outcome negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	True positive rate (TPR), Sensitivity, Recall $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$

ROC curve



- http://www.navan.name/roc/

Note: ROC curve needs a probability

Picking a good performance metrics is still a open question

The paper Facing Imbalanced Data Recommendations for the Use of Performance Metrics found that "while ROC was unaffected by skew, the precision-recall curves suggest that ROC may mask poor performance in some cases."

Picking a good performance metrics is still a open question, but in publications, competitions etc ROC is a safe bet.

Always check what your problem needs first

All in all, things to keep in mind

The measure you optimize to makes a difference

The measure you report makes a difference

Use measure appropriate for problem/community

Accuracy often is not sufficient/appropriate;

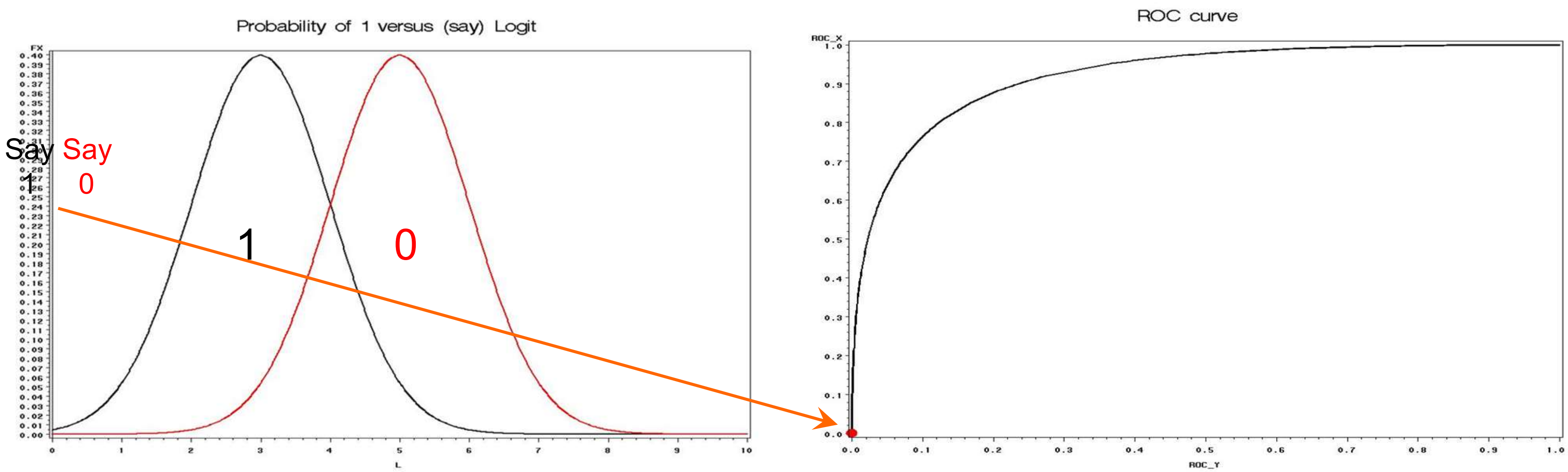
ROC is gaining popularity in the ML community

Only accuracy generalizes to >2 classes!

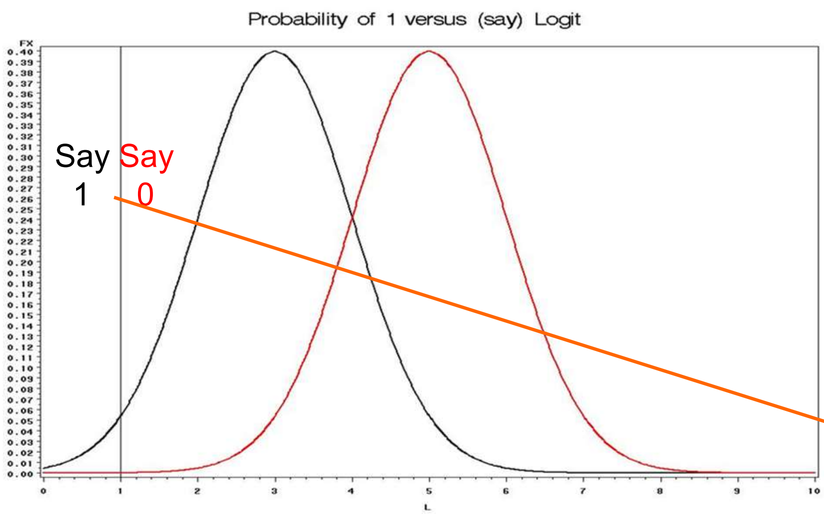
END

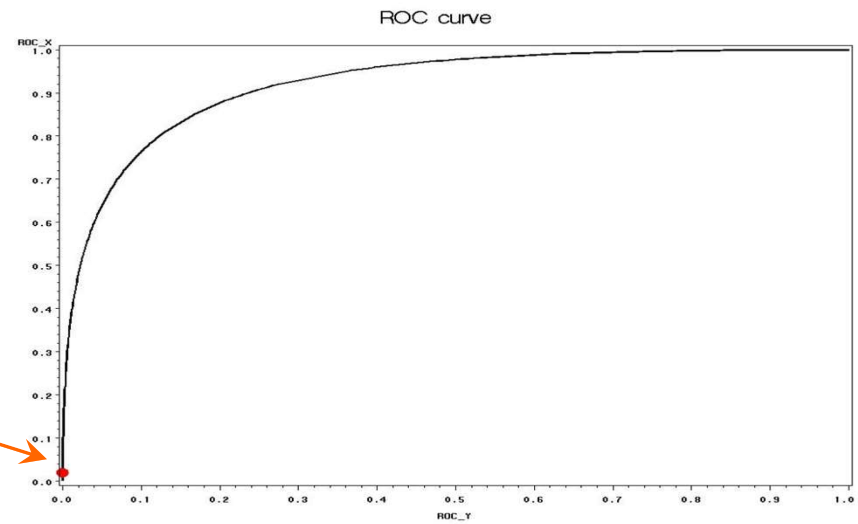
ROC Curve Demo

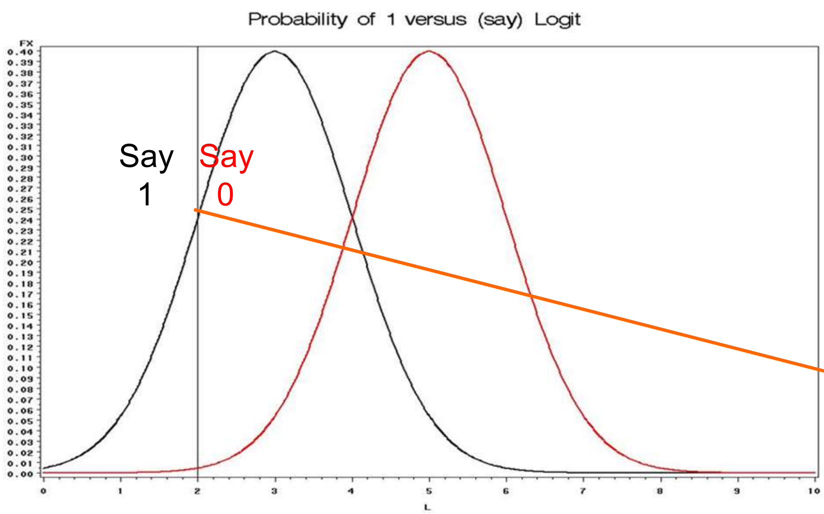
From Dave Dickey, used with permission


Probability of 1 versus (say) Logit
FX
L
Say
1
Say
0
1
0
ROC curve
ROC_X
ROC_Y

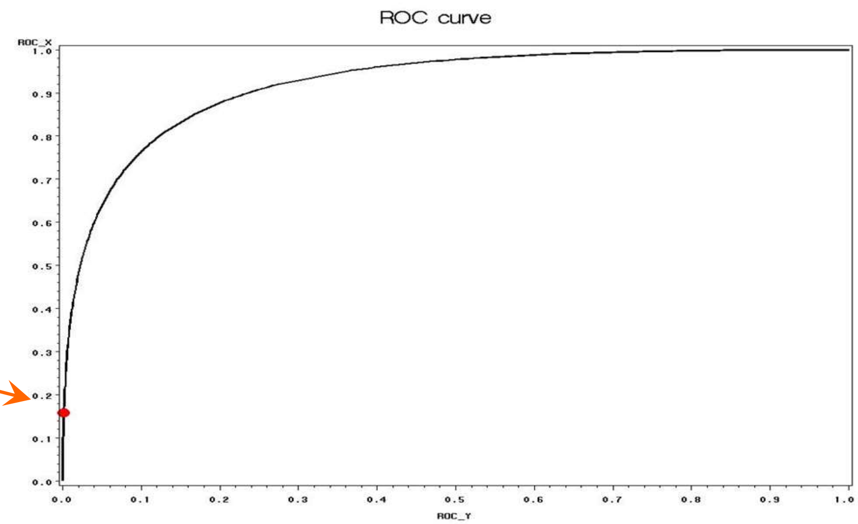
%0s


Probability of 1 versus (say) Logit
FX
L
Say 1
Say 0

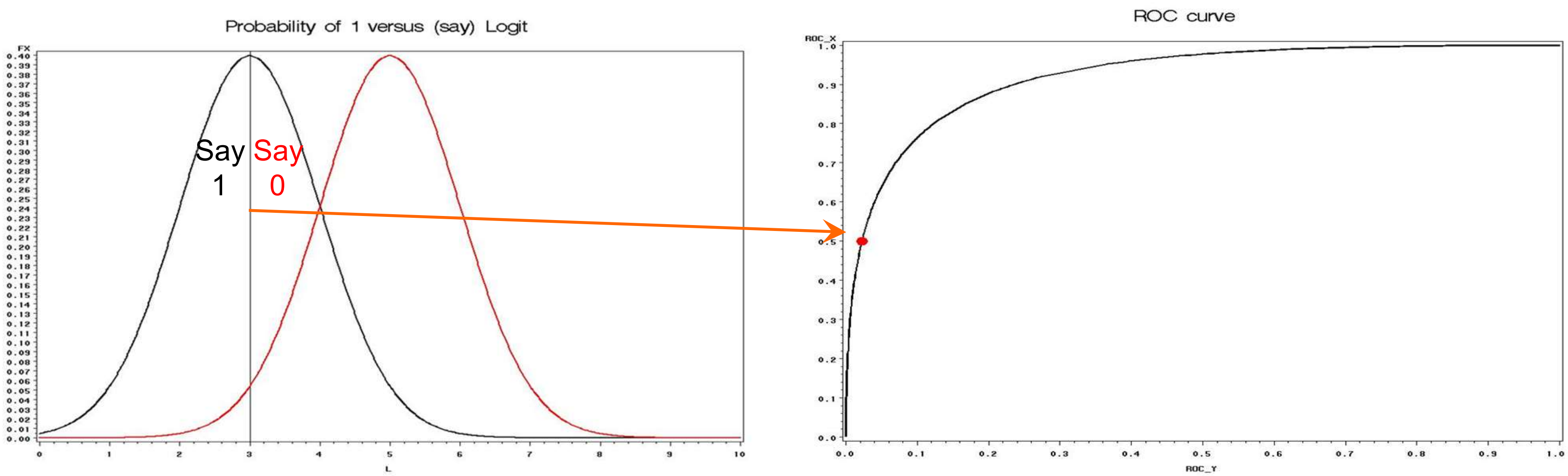

ROC curve
ROC_X
ROC_Y

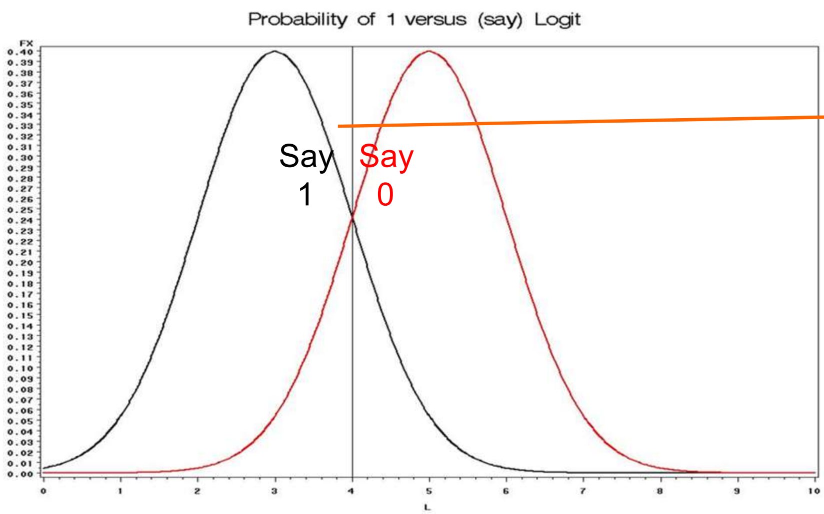


Probability of 1 versus (say) Logit
FX
L
Say 1
Say 0

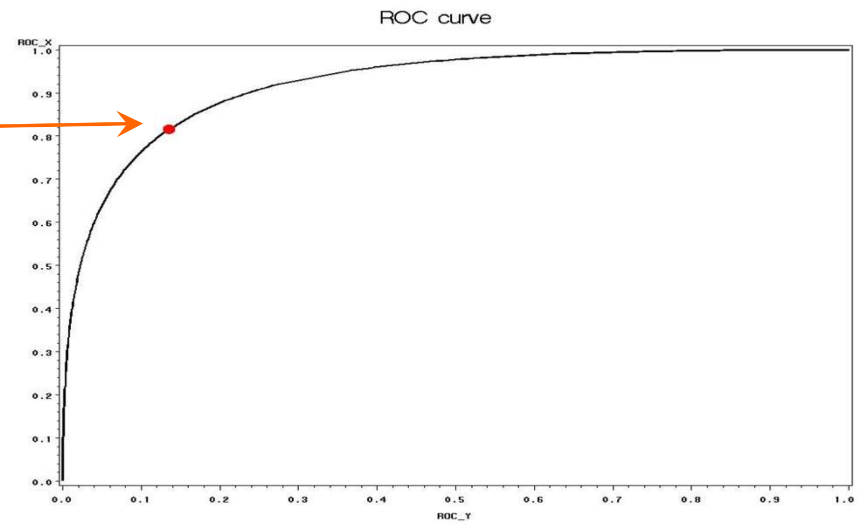


ROC curve
ROC_X
ROC_Y

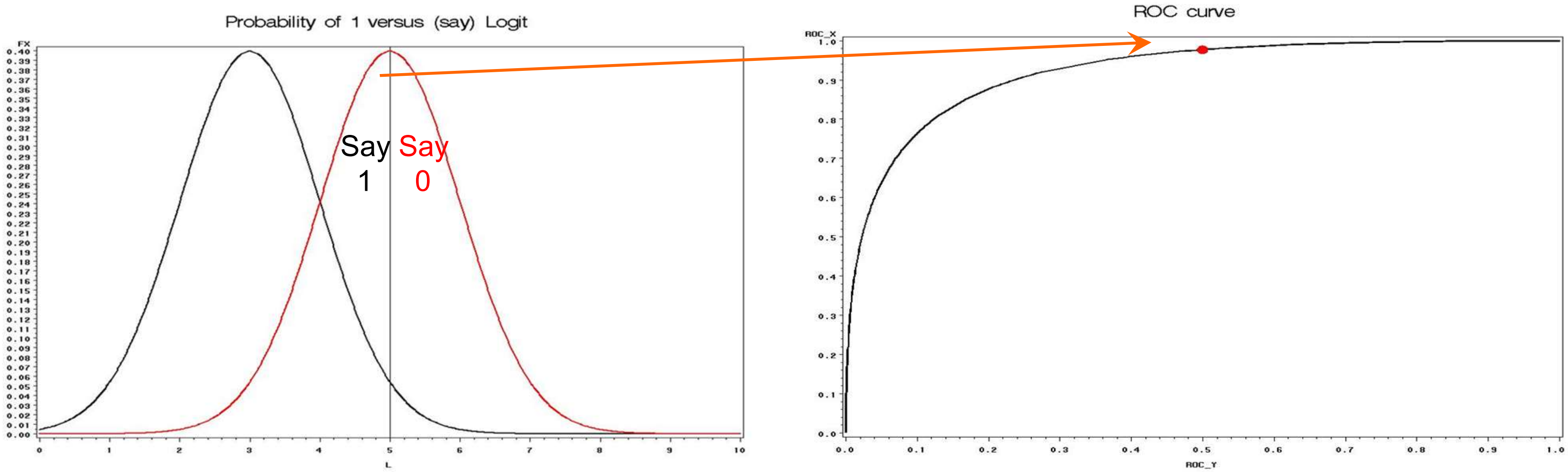

Probability of 1 versus (say) Logit
FX
L
Say 1
Say 0
ROC curve
ROC_X
ROC_Y

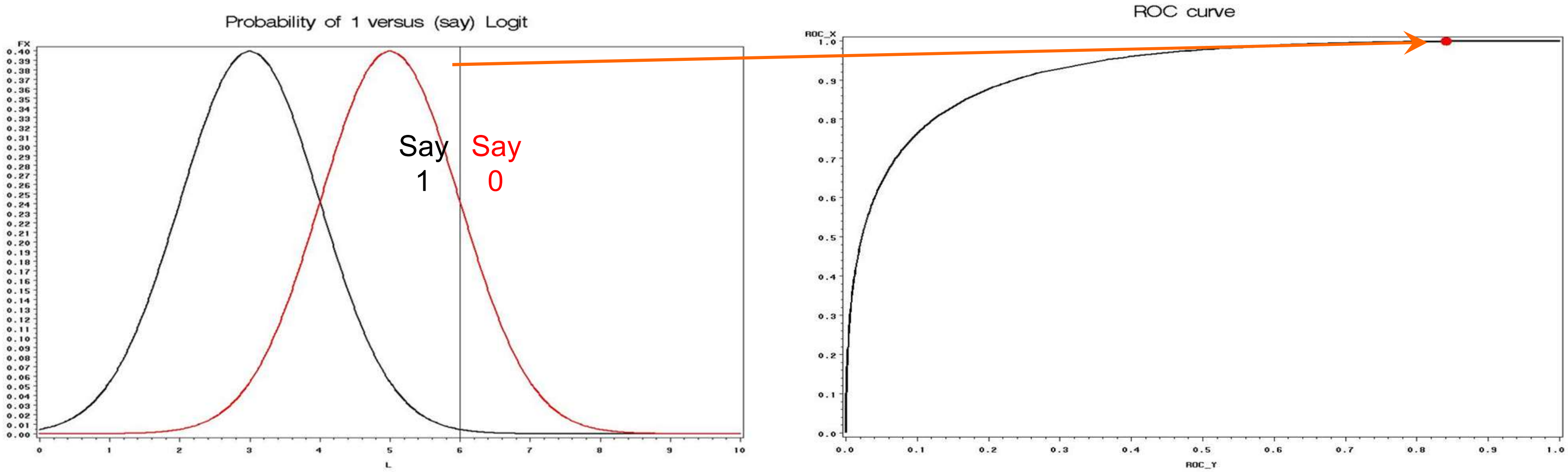


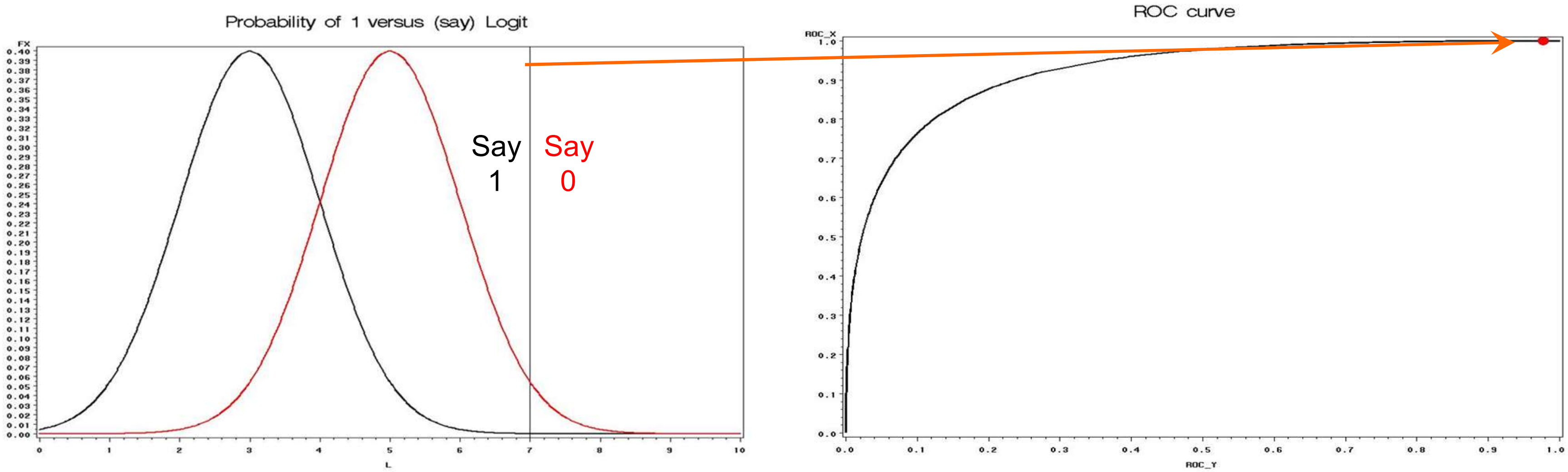
Probability of 1 versus (say) Logit
FX
L
Say 1
Say 0



ROC curve
ROC_X
ROC_Y


Probability of 1 versus (say) Logit
FX
L
Say
1
Say
0
ROC curve
ROC_X
ROC_Y


Probability of 1 versus (say) Logit
FX
Say
1
Say
0
L
ROC curve
ROC_X
ROC_Y

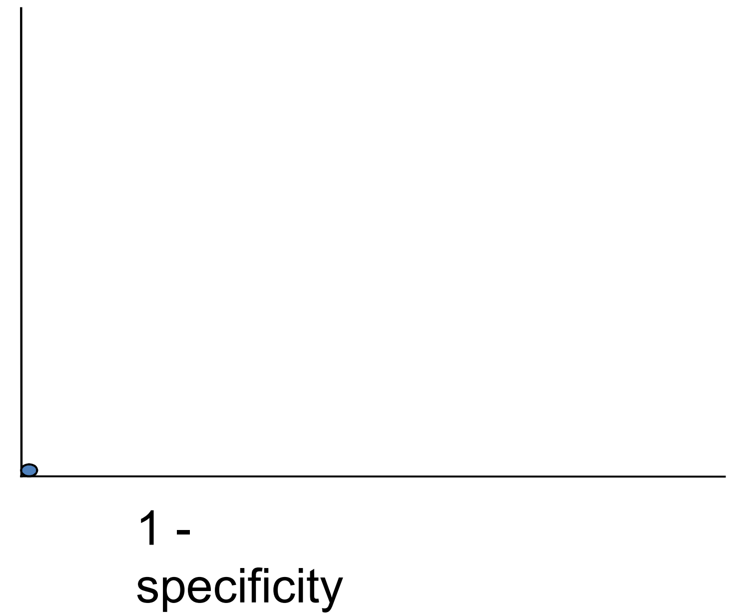

Probability of 1 versus (say) Logit
FX
L
Say
1
Say
0
ROC curve
ROC_X
ROC_Y

Example 2

	0	1
X=10	2	18
	8	12
X=13	15	5
X=16		

Say 1

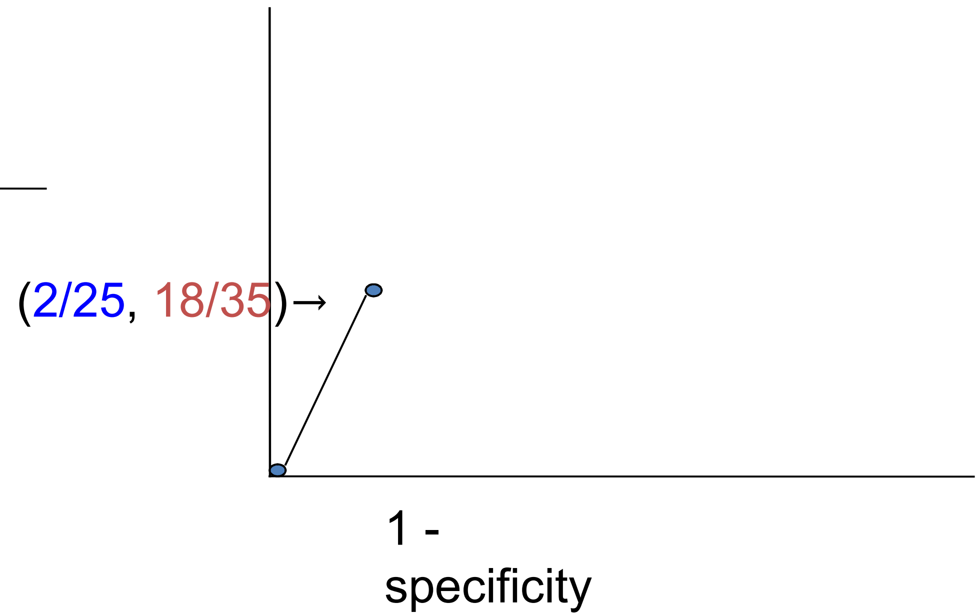
Say 0



Example 2

	0	1
X=10	2	18
	8	12
X=13	15	5
X=16		

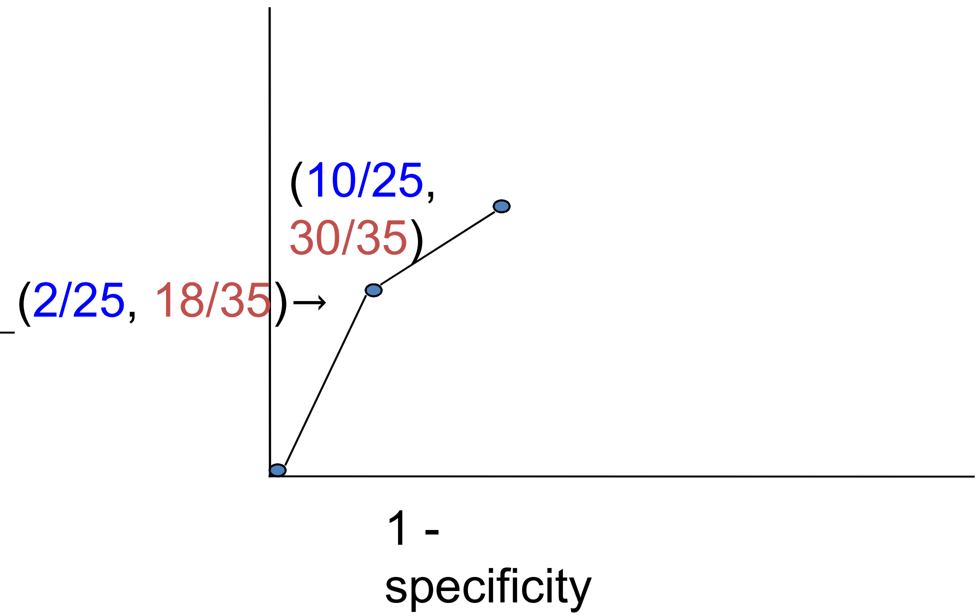
Say 1
Say 0



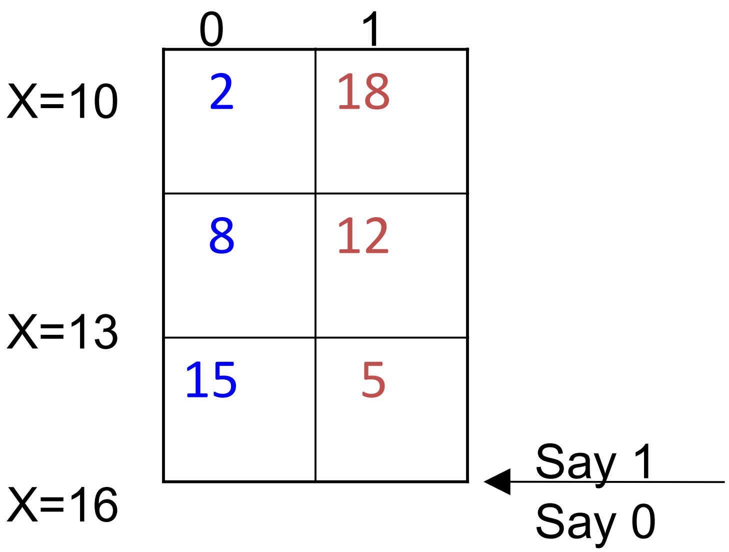
Example 2

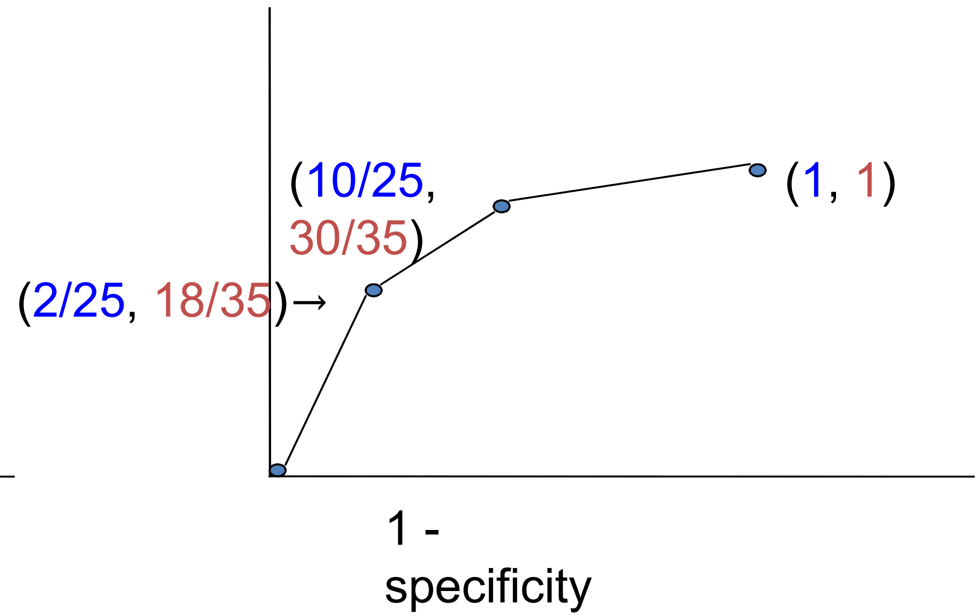
	0	1
X=10	2	18
	8	12
X=13	15	5
X=16		

Say 1
Say 0

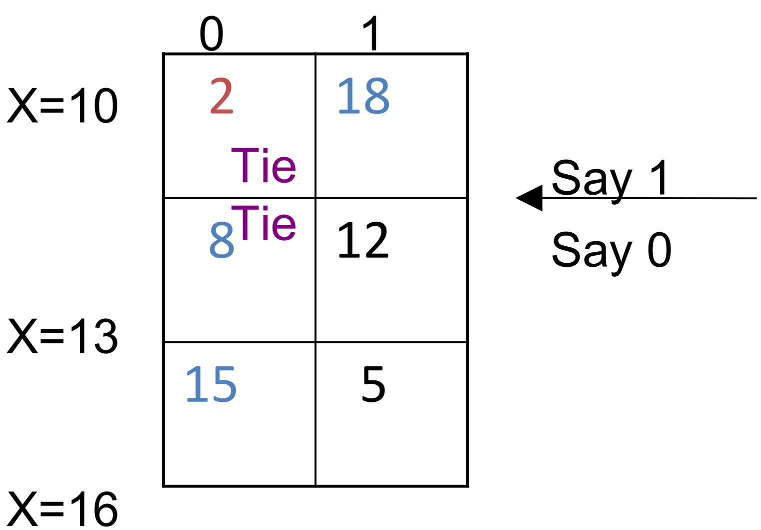


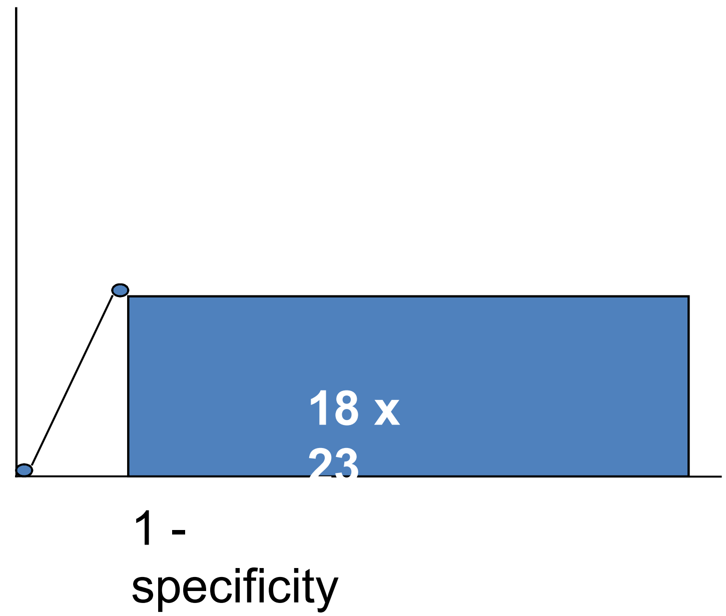
Example 2


0
1
X=10
2
18
8
12
X=13
15
5
X=16
Say 1
Say 0

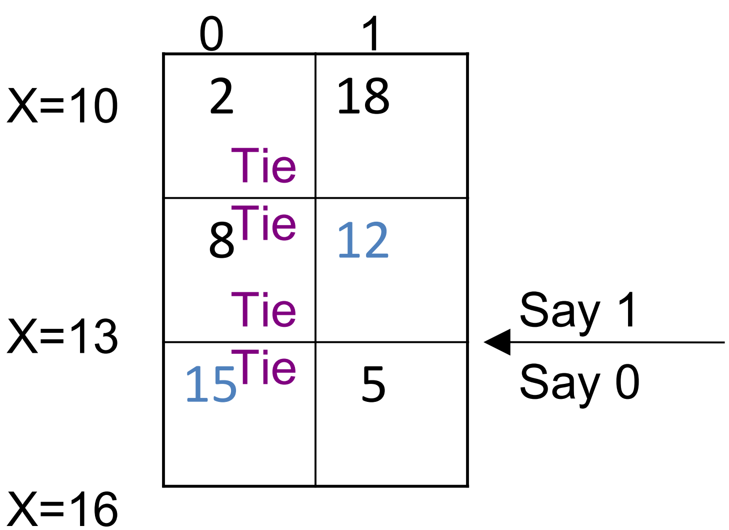

(10/25, 30/35)
(1, 1)
(2/25, 18/35)→
1 - specificity

Example 2

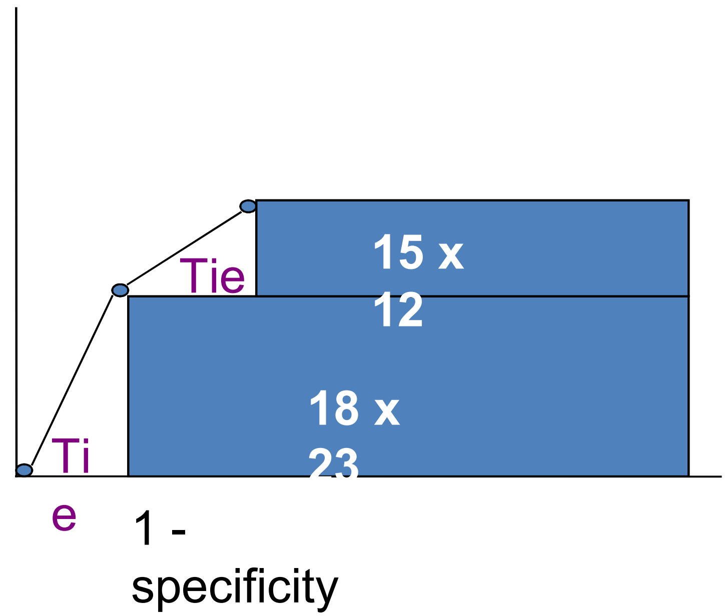

0
1
X=10
2
18
Tie
Tie
8
12
X=13
15
5
X=16
Say 1
Say 0


18 x
23
1 -
specificity

Example 2



0
1
X=10
2
18
Tie
Tie
8
12
Tie
X=13
15
Tie
5
X=16
Say 1
Say 0

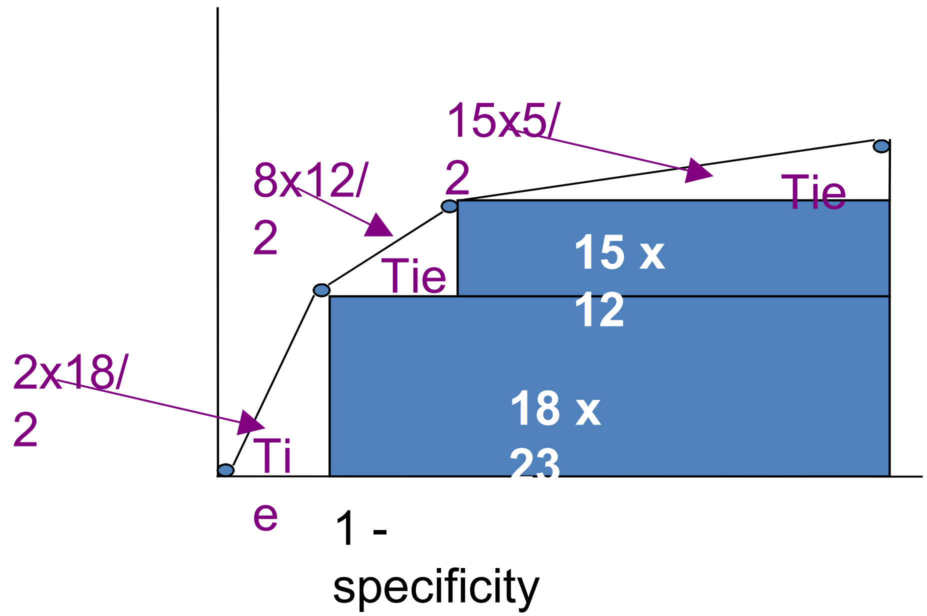


15 x
12
Tie
18 x
23
Ti
e
1 -
specificity

Example 2

	0	1
X=10	2 Tie	18
	8 Tie Tie	12
X=13	15 Tie Tie	5
X=16	Tie	

Say 1
Say 0



AUC Area Under Curve = concordant pair plus ½ ties (proportions)

Pairs: 25x35=875 with 0 paired with 1.

Proportion Concordant:

(18x23+12x15)/(25x35)=(18/25)x(23/35)+(12/25)x(15/35)

½ Proportion Ties ½(2x18 + 8x12 + 15x5)/(25x35)