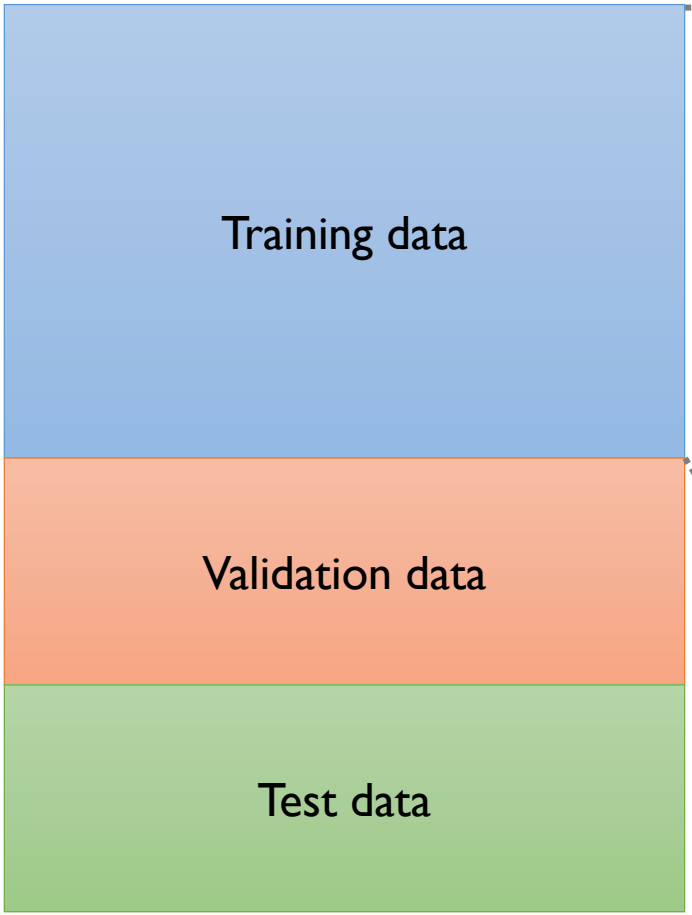
Classification Performance Evaluation

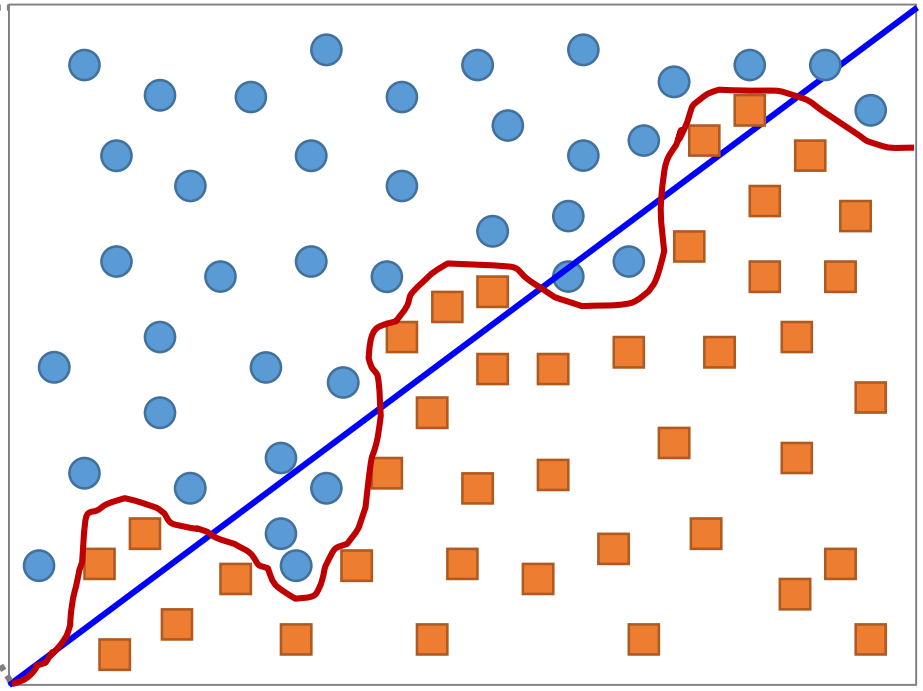
강필성
고려대학교 산업경영공학부
pilsung_kang@korea.ac.kr

AGENDA

Why Evaluate?

- Over-fitting for training data

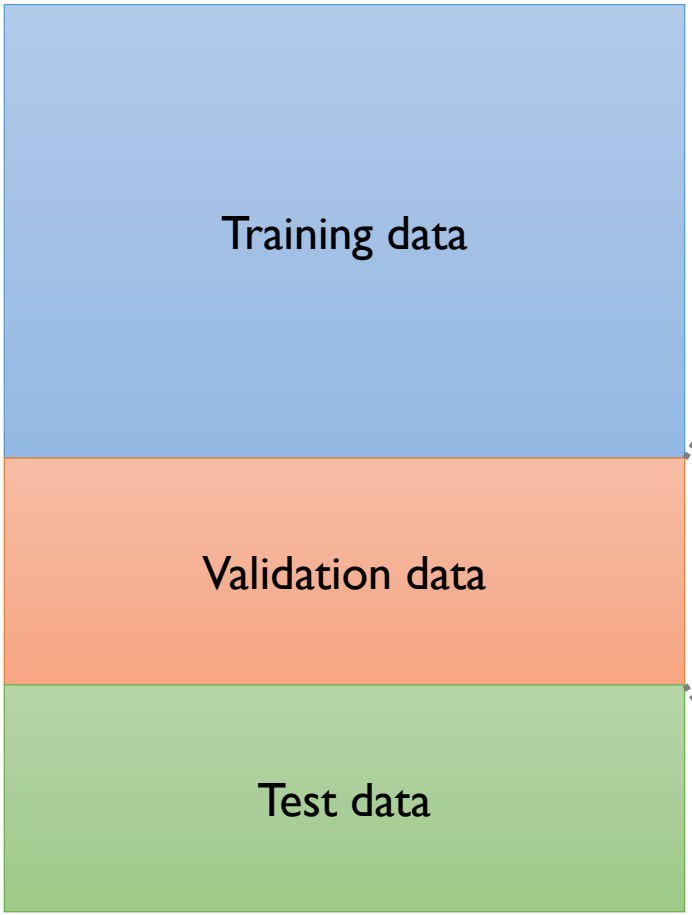




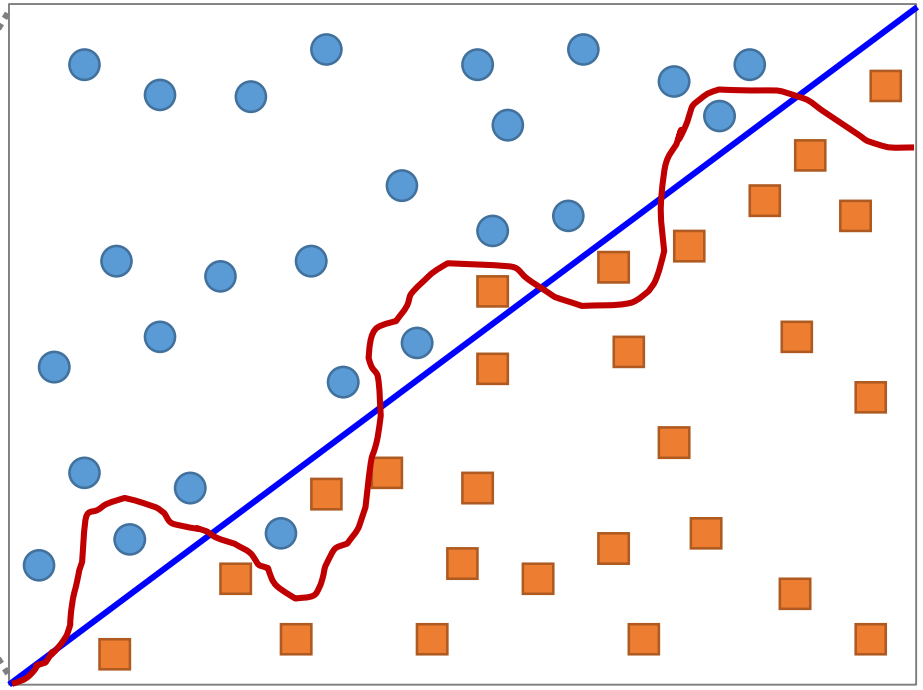
Is red boundary better than blue one?

Why Evaluate?

- Over-fitting for training data



Do not memorize them all!!



Why Evaluate?

- Multiple methods are available to classify or predict.
  - ✓ Classification:
    - Naïve bayes, linear discriminant, k-nearest neighbor, classification trees, etc.
  - ✓ Prediction:
    - Multiple linear regression, neural networks, regression trees, etc.
- For each method, multiple choices are available for settings.
  - ✓ Neural networks: # hidden nodes, activation functions, etc.
- To choose best model, need to assess each model's performance.
  - ✓ Best setting (parameters) among various candidates for an algorithm (validation).
  - ✓ Best model among various data mining algorithms for the task (test).

Classification Performance

Example: Gender classification

- Classify a person based on his/her body fat percentage (BFP).

10.0	21.7	8.9	19.9	23.4	28.9	15.7	21.6	21.5	23.2

- Simple classifier: if BFP > 20 then female else male.

10.0	21.7	8.9	19.9	23.4	28.9	15.7	21.6	21.8	23.2
M	F	M	M	F	F	M	F	F	F

- How do you evaluate the performance of the above classifier?

Classification Performance

2

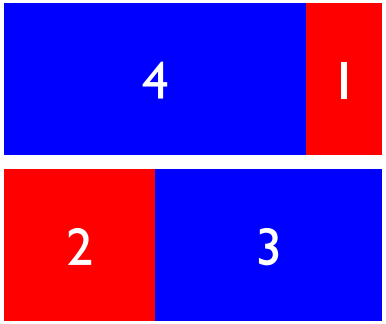
Confusion Matrix

- Summarizes the correct and incorrect classifications that a classifier produced for a certain data set.

10.0	21.7	8.9	19.9	23.4	28.9	15.7	21.6	21.5	23.2
M	F	M	M	F	F	M	F	F	F

- Confusion matrix can be constructed as

Confusion Matrix		Predicted	
		F	M
Actual	F	4	1
	M	2	3



Classification Performance

2

Confusion Matrix

- Summarizes the correct and incorrect classifications that a classifier produced for a certain data set.

Confusion Matrix		Predicted	
		1(+)	0(-)
Actual	1(+)	n_{11}	n_{10}
	0(-)	n_{01}	n_{00}

Confusion Matrix		Predicted	
		F	M
Actual	F	4	1
	M	2	3

- Misclassification error = $(n_{01}+n_{10})/(n_{11}+n_{10}+n_{01}+n_{00}) = (2+1)/10 = 0.3$
- Accuracy = (1-Misclassification error) = $(n_{11}+n_{00})/(n_{11}+n_{10}+n_{01}+n_{00}) = (4+3)/10 = 0.7$

Classification Performance

2

Confusion Matrix

- Summarizes the correct and incorrect classifications that a classifier produced for a certain data set.

Confusion Matrix		Predicted	
		1(+)	0(-)
Actual	1(+)	n_{11}	n_{10}
	0(-)	n_{01}	n_{00}

Confusion Matrix		Predicted	
		F	M
Actual	F	4	1
	M	2	3

- Balanced correction rate (BCR): $\sqrt{\frac{n_{11}}{n_{11}+n_{10}} \cdot \frac{n_{00}}{n_{01}+n_{00}}} = \sqrt{0.8 \times 0.6} = 0.69$

- F1-Measure: $\frac{2 \times Recall \times Precision}{Recall + Precision} = \frac{2 \times 0.8 \times 0.67}{0.8 + 0.67} = 0.85$

Classification Performance

3

Cut-off for classification

- A new classifier: : if BFP > θ then female else male.

10.0	23.6	8.9	19.9	25.4	28.6	15.7	22.7	21.5	24.2

- Sort data in a descending order of BFS.

28.6	25.4	24.2	23.6	22.7	21.5	19.9	15.7	10.0	8.9

- How do you decide the cut-off for classification?

Classification Performance

3

Cut-off for classification

- Performance measures for different cut-offs:

No.	BFS	Gender
1	28.6	F
2	25.4	M
3	24.2	F
4	23.6	F
5	22.7	F
6	21.5	M
7	19.9	F
8	15.7	M
9	10.0	M
10	8.9	M

- If $\theta = 24$,

Confusion Matrix		Predicted	
		F	M
Actual	F	2	3
	M	1	4

- Misclassification error: 0.4
- Accuracy: 0.6
- Balanced correction rate: 0.57
- F1 measure = 0.5

Classification Performance

3

Cut-off for classification

- Performance measures for different cut-offs:

No.	BFS	Gender
1	28.6	F
2	25.4	M
3	24.2	F
4	23.6	F
5	22.7	F
6	21.5	M
7	19.9	F
8	15.7	M
9	10.0	M
10	8.9	M

- If $\theta = 22$,

Confusion Matrix		Predicted	
		F	M
Actual	F	4	1
	M	1	4

- Misclassification error: 0.2
- Accuracy: 0.8
- Balanced correction rate: 0.8
- F1 measure = 0.8

Classification Performance

3

Cut-off for classification

- Performance measures for different cut-offs:

No.	BFS	Gender
1	28.6	F
2	25.4	M
3	24.2	F
4	23.6	F
5	22.7	F
6	21.5	M
7	19.9	F
8	15.7	M
9	10.0	M
10	8.9	M

- If $\theta = 18$,

Confusion Matrix		Predicted	
		F	M
Actual	F	5	0
	M	2	3

- Misclassification error: 0.2
- Accuracy: 0.8
- Balanced correction rate: 0.77
- F1 measure = 0.83

Classification Performance

Cut-off for classification

- In general, classification algorithms can produce the likelihood for each class in terms of probability or degree of evidence, etc.

3

- Classification performance highly depends on the cut-off of the algorithm.

- For model selection & model comparison, cut-off independent performance measures are recommended.

- Lift charts, receiver operating characteristic (ROC) curve, etc.

Classification Performance

- Area Under Receiver Operating Characteristic Curve (AUROC)
  - ✓ Fault Detection Problem:
    - Classify Good/Faulty products
    - A total of 100 products
    - 20 products are fault (Fault ratio: 0.2)
    - Label: 1(NG), 0(G)

Classification Performance

- Estimated likelihood (P(NG)) and the target label information

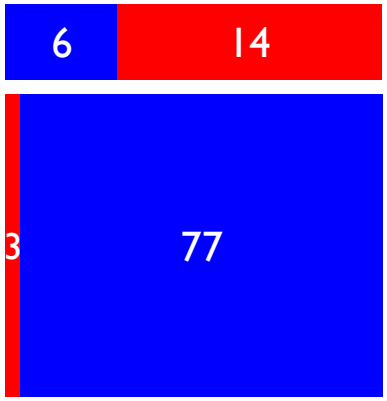
Glass	P(NG)	Label	Glass	P(NG)	Label	Glass	P(NG)	Label	Glass	P(NG)	Label
1	0.976	1	26	0.716	1	51	0.41	0	76	0.186	0
2	0.973	1	27	0.676	0	52	0.406	1	77	0.183	0
3	0.971	0	28	0.672	0	53	0.378	0	78	0.178	0
4	0.967	1	29	0.662	0	54	0.376	0	79	0.176	0
5	0.937	0	30	0.647	0	55	0.362	0	80	0.173	0
6	0.936	1	31	0.64	1	56	0.355	0	81	0.17	0
7	0.929	1	32	0.625	0	57	0.343	0	82	0.133	0
8	0.927	0	33	0.624	0	58	0.338	0	83	0.12	0
9	0.923	1	34	0.613	1	59	0.335	0	84	0.119	0
10	0.898	0	35	0.606	0	60	0.334	0	85	0.112	0
11	0.863	1	36	0.604	0	61	0.328	0	86	0.093	0
12	0.862	1	37	0.601	0	62	0.313	0	87	0.086	0
13	0.859	0	38	0.594	0	63	0.285	1	88	0.079	0
14	0.855	0	39	0.578	0	64	0.274	0	89	s0.071	0
15	0.847	1	40	0.548	0	65	0.273	0	90	0.069	0
16	0.845	1	41	0.539	1	66	0.272	0	91	0.047	0
17	0.837	0	42	0.525	1	67	0.267	0	92	0.029	0
18	0.833	0	43	0.524	0	68	0.265	0	93	0.028	0
19	0.814	0	44	0.514	0	69	0.237	0	94	0.027	0
20	0.813	0	45	0.51	0	70	0.217	0	95	0.022	0
21	0.793	1	46	0.509	0	71	0.213	0	96	0.019	0
22	0.787	0	47	0.455	0	72	0.204	1	97	0.015	0
23	0.757	1	48	0.449	0	73	0.201	0	98	0.01	0
24	0.741	0	49	0.434	0	74	0.2	0	99	0.005	0
25	0.737	0	50	0.414	0	75	0.193	0	100	0.002	0

Classification Performance

Confusion matrix

- Set the cut-off to 0.9
  - Malignant if P(Malignant) > 0.9, else benign.

Confusion Matrix		Predicted	
		M	B
Actual	M	6	14
	B	3	77



  - Misclassification error = 0.17
  - Accuracy = 0.83
- Is it a good classification model?

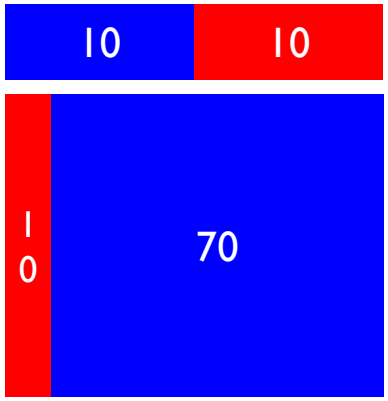
4

Classification Performance

Confusion matrix

- Set the cut-off to 0.8
  - Malignant if P(Malignant) > 0.8, else benign.

Confusion Matrix		Predicted	
		M	B
Actual	M	10	10
	B	10	70



  - Misclassification error = 0.2
  - Accuracy = 0.8

- Is it worse than the previous model?

4

Classification Performance

Receiver operating characteristics (ROC) curve

- Sort the records based on the P(interesting class) in a descending order.
- Compute the true positive rate and false positive rate by varying the cut-off.
- Draw a chart where x & y axes are false & true positive, respectively.

5

Classification Performance

ROC example

- First cut-off

Glass	P(NG)	Label
1	0.976	1
2	0.973	1
3	0.971	0
4	0.967	1
5	0.937	0
⋮	⋮	⋮

Confusion Matrix		예측	
		NG	G
실제	NG	0	20
	G	0	80

$$\text{TPR} = \frac{0}{20} = 0$$

$$\text{FPR} = \frac{0}{80} = 0$$

5

Classification Performance

ROC example

- Second cut-off

Glass	P(NG)	Label	TPR	FPR
			0	0
1	0.976	1		
2	0.973	1		
3	0.971	0		
4	0.967	1		
5	0.937	0		
⋮	⋮	⋮	⋮	⋮

Confusion Matrix		예측	
		NG	G
실제	NG	1	19
	G	0	80

$$\text{TPR} = \frac{1}{20} = 0.05$$

$$\text{FPR} = \frac{0}{80} = 0$$

5

Classification Performance

ROC example

- Third cut-off

Glass	P(NG)	Label	TPR	FPR
			0	0
1	0.976	1	0.05	0
2	0.973	1		
3	0.971	0		
4	0.967	1		
5	0.937	0		
⋮	⋮	⋮	⋮	⋮

Confusion Matrix		예측	
		NG	G
실제	NG	2	18
	G	0	80

$$\text{TPR} = \frac{2}{20} = 0.10$$

$$\text{FPR} = \frac{0}{80} = 0$$

5

Classification Performance

ROC example

- Fourth cut-off

Glass	P(NG)	Label	TPR	FPR
			0.00	0.00
1	0.976	1	0.05	0.00
2	0.973	1	0.10	0.00
3	0.971	0		
4	0.967	1		
5	0.937	0		
⋮	⋮	⋮	⋮	⋮

Confusion Matrix		예측	
		NG	G
실제	NG	2	18
	G	1	79

$$\text{TPR} = \frac{2}{20} = 0.10$$

$$\text{FPR} = \frac{1}{80} = 0.0125$$

5

Classification Performance

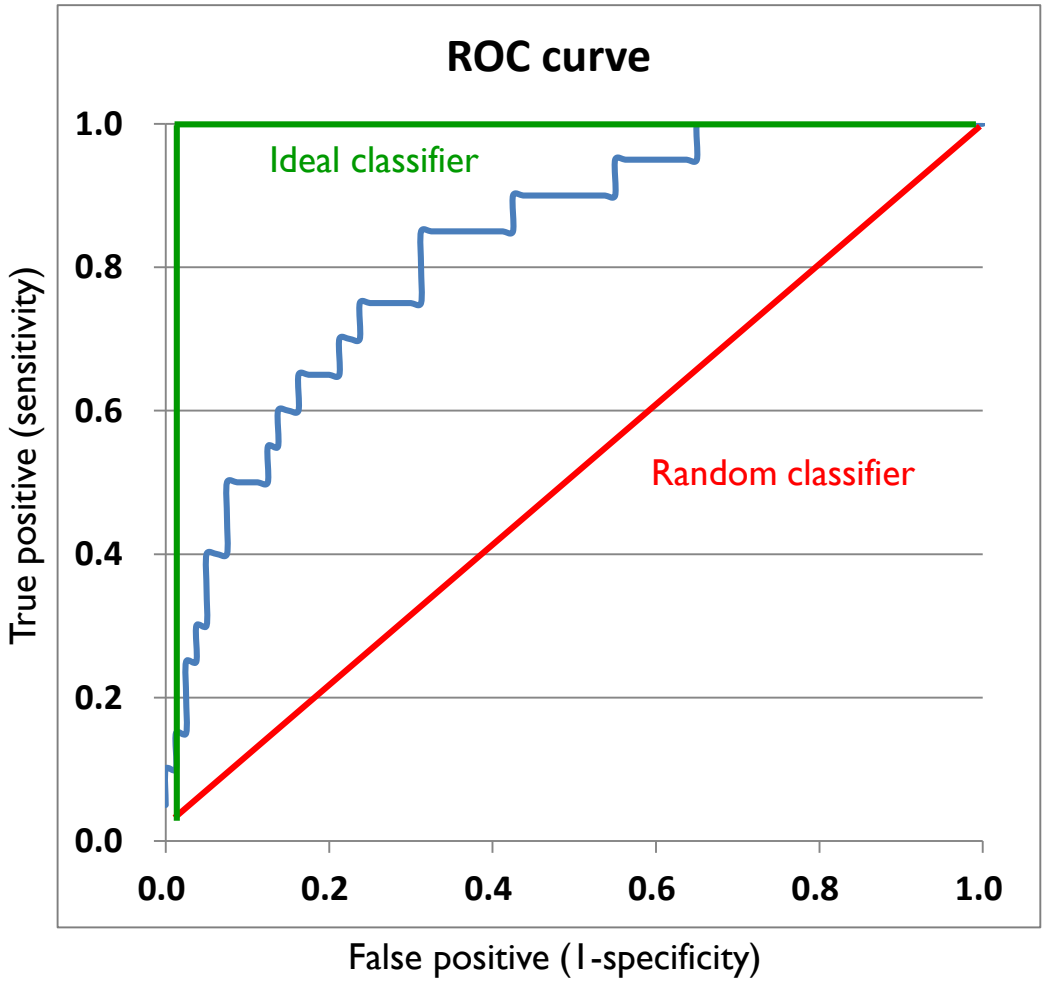
ROC example

- Compute all possible TPR and FPR
- Draw a graph with FPR as an x-axis and TPR as an y-axis

Glass	P(NG)	Label	TPR	FPR
			0.000	0.000
1	0.976	1	0.050	0.000
2	0.973	1	0.100	0.000
3	0.971	0	0.100	0.013
4	0.967	1	0.150	0.013
5	0.937	0	0.150	0.025
6	0.936	1	0.200	0.025
7	0.929	1	0.250	0.025
8	0.927	0	0.250	0.038
⋮	⋮	⋮	⋮	⋮
96	0.019	0	1.000	0.950
97	0.015	0	1.000	0.963
98	0.01	0	1.000	0.975
99	0.005	0	1.000	0.988
100	0.002	0	1.000	1.000

Classification Performance

Receiver operating characteristics (ROC) curve

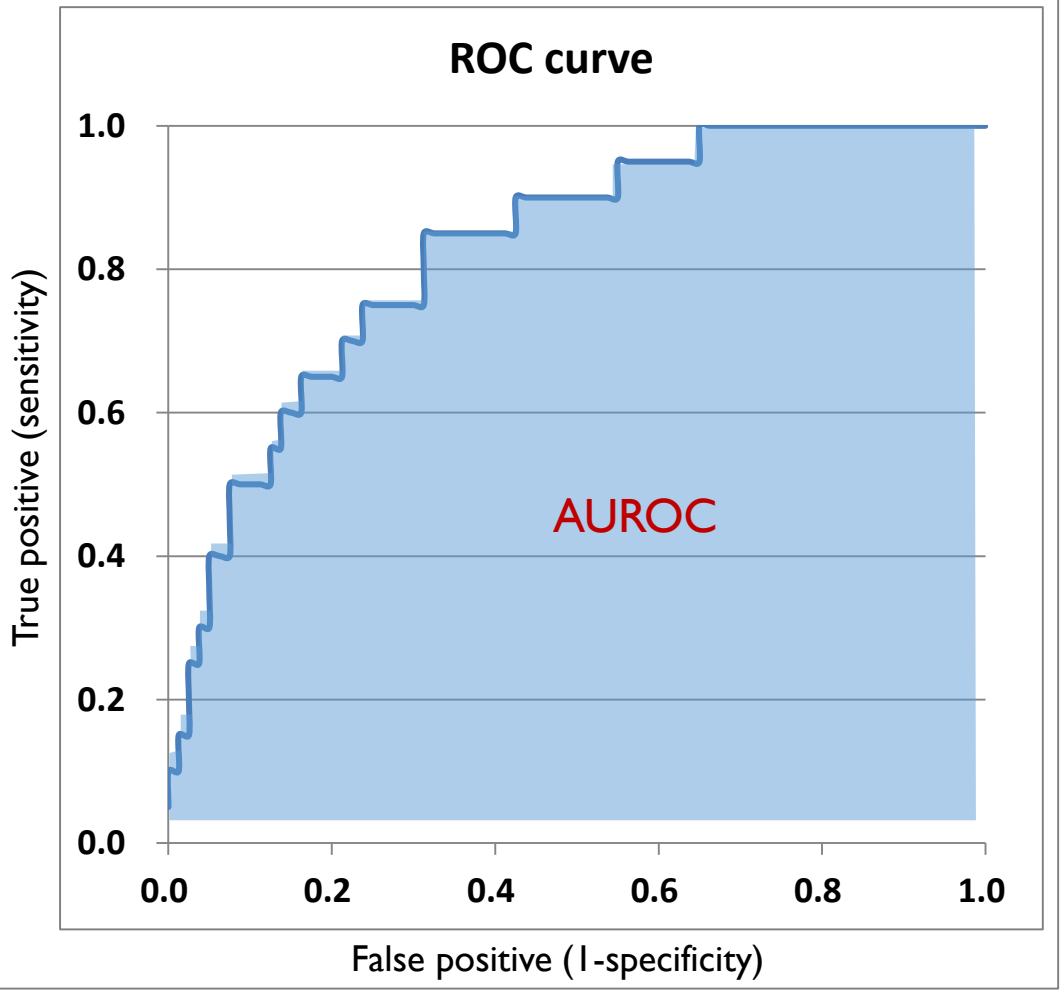


5

Classification Performance

Area Under ROC curve (AUROC)

- The area under the ROC curve.
- Can be a useful metric for parameter/model selection.
- 1 for the ideal classifier
- 0.5 for the random classifier.



6


ANY
questions?