11.01.2019

Statistical Methods in AI (CSE/ECE 471)

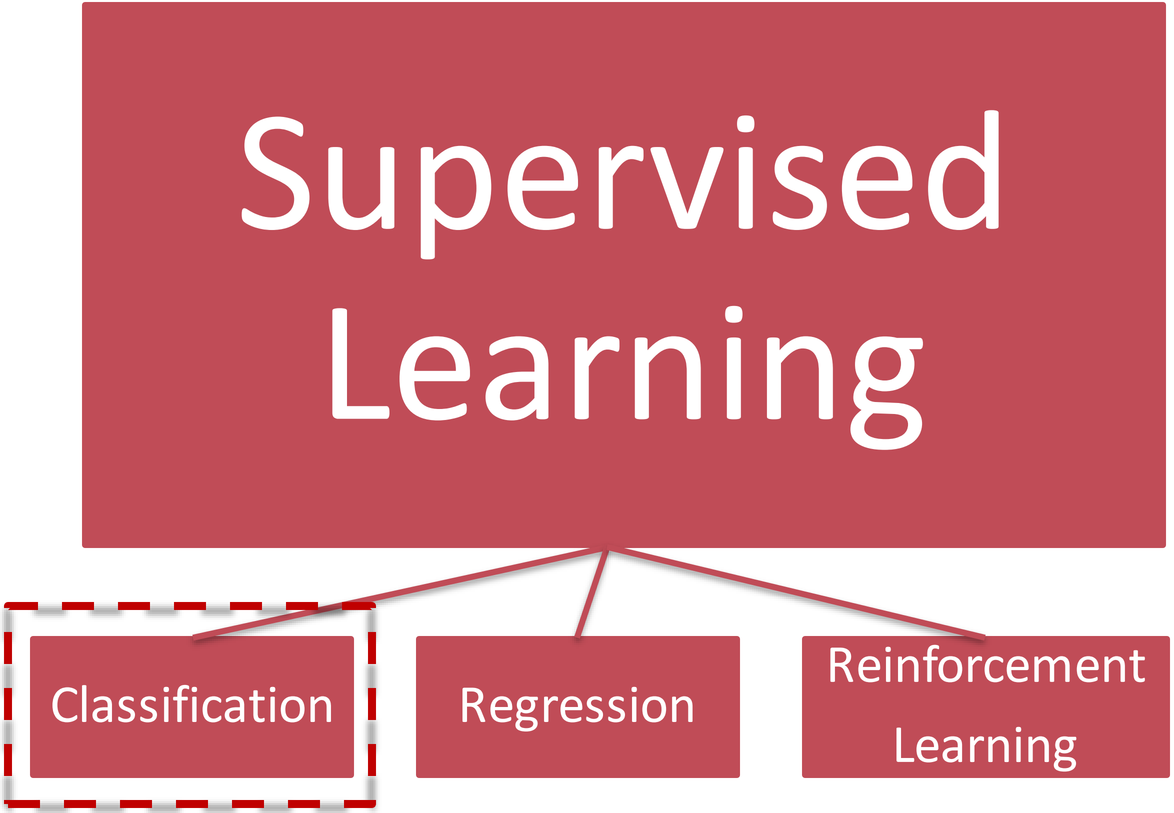
Lecture-4: Intro to Performance Measures, Benchmarking

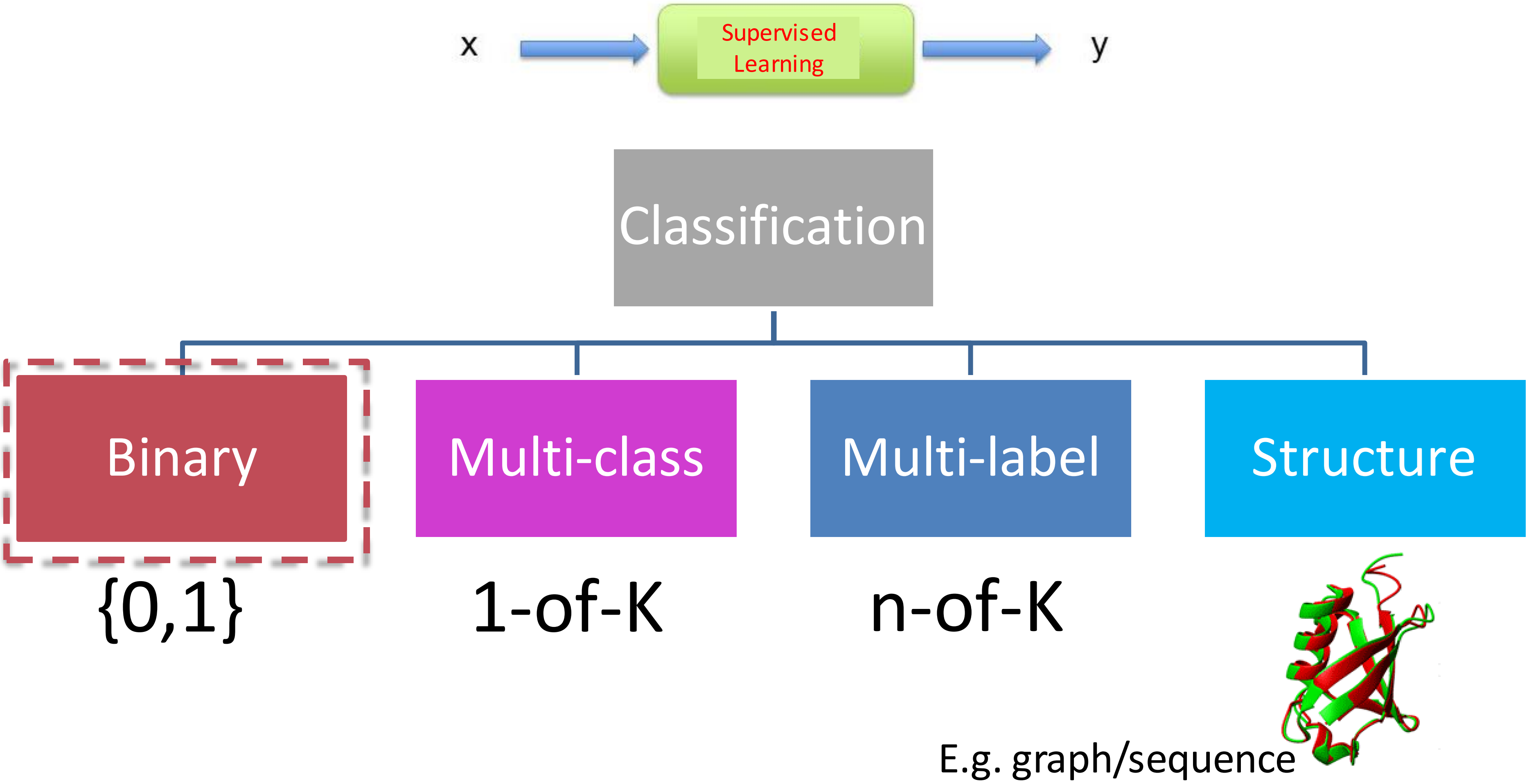
Ravi Kiran

Center for Visual Information Technology (CVIT), IIIT Hyderabad

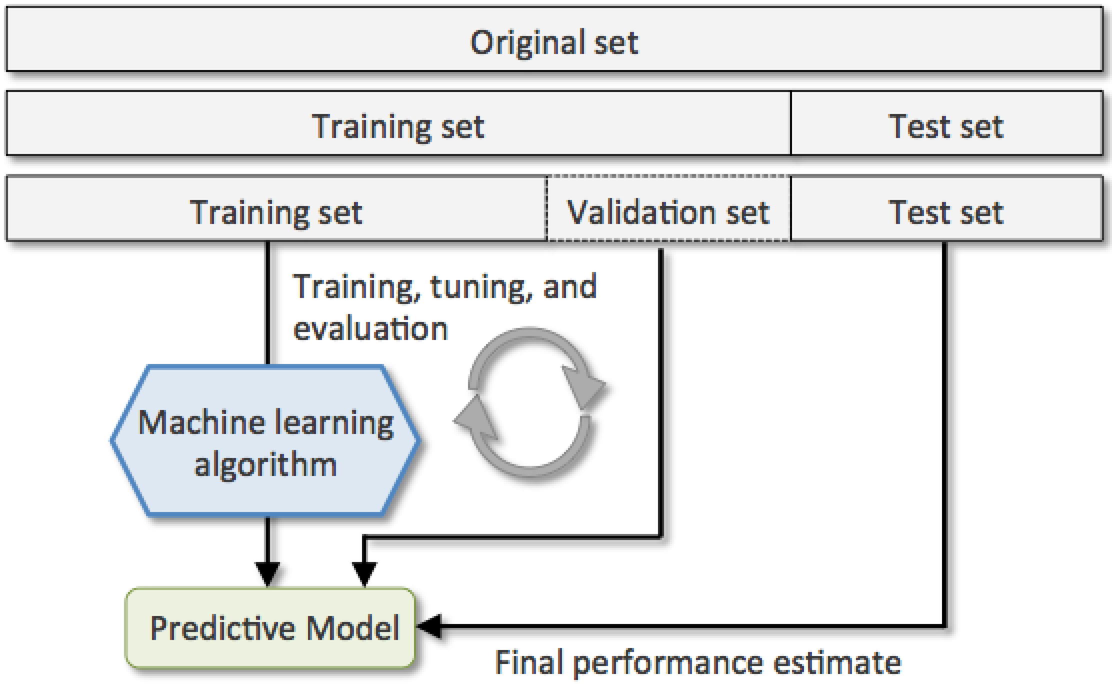
Announcements

- A1 has been posted. Due: **20/1, 11.59 PM**
- This week's tutorial: Probability recap, ML datasets, visualization approaches. **Bring your laptops**.

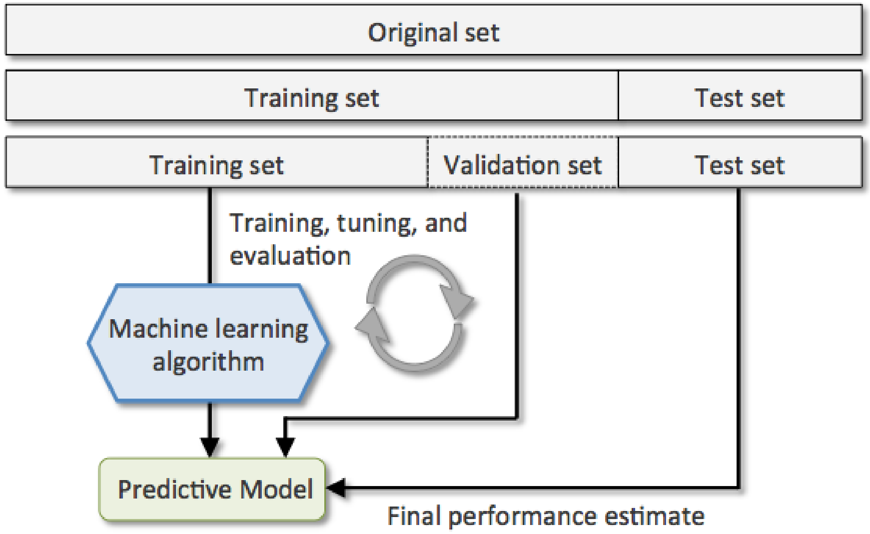

Supervised Learning
Classification
Regression
Reinforcement Learning

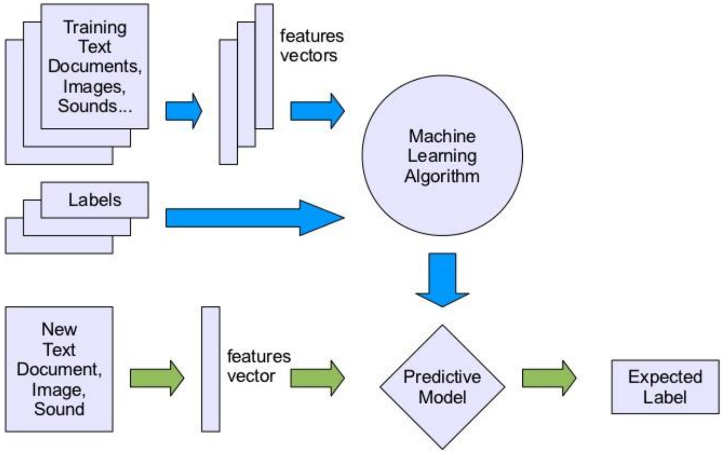

x
Supervised Learning
y
Classification
Binary
Multi-class
Multi-label
Structure
{0,1}
1-of-K
n-of-K
E.g. graph/sequence

The Train-Validation-Test paradigm

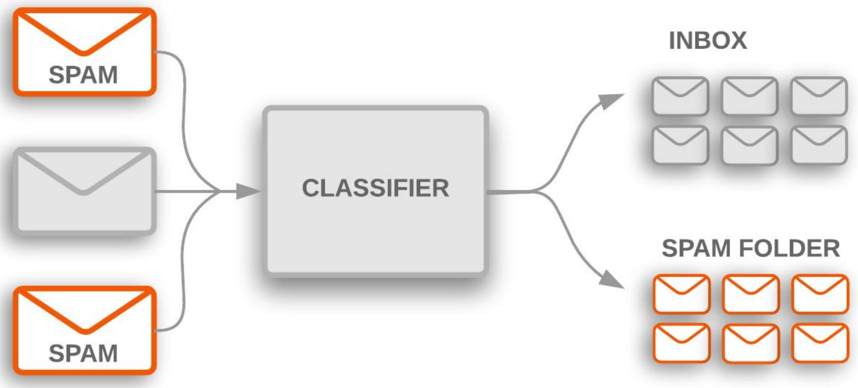


The Train-Validation-Test paradigm



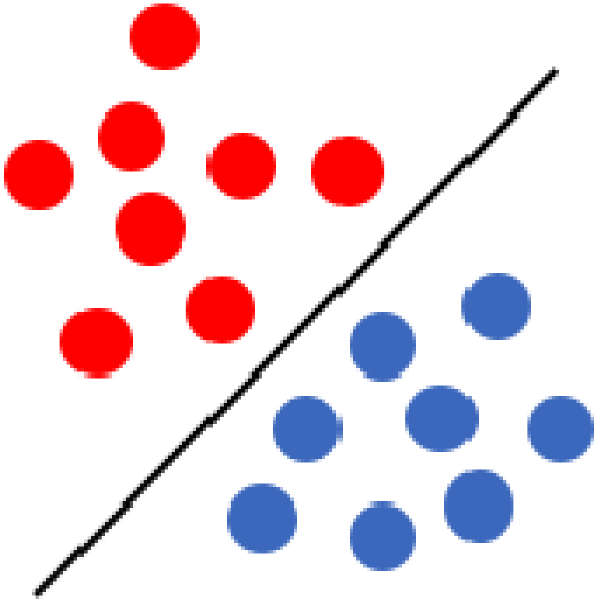


Binary Classification



Binary classification

Positive → E.g. Spam = YES



Negative → E.g. Spam = NO

Binary case...

$$Accuracy = \frac{(100 + 50)}{165} = 0.91$$

$$Misclassification = \frac{(10 + 5)}{165} = 0.09$$

$$TruePositiveRate(TP) = \frac{(100)}{105} = 0.95$$

$$FalsePositiveRate(FP) = \frac{(10)}{60} = 0.17$$

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Binary case...

$$TrueNegativeRate(TN) = \frac{(50)}{60} = 0.833$$

$$FalseNegativeRate(FN) = \frac{(5)}{105} = 0.048$$

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Key accuracy measures and terminologies

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

- Classification Error = $\frac{errors}{total}$

  = $\frac{FP + FN}{TP + TN + FP + FN}$

- Accuracy = 1 - Error = $\frac{correct}{Total}$

  = $\frac{TP + TN}{TP + TN + FP + FN}$

Precision and Recall

- **Cancer-Prediction System**
- **Pool of 100 patients' data**
- 3 patients' data from the pool are selected for chemotherapy ; Rest (100-3=97) are declared healthy !
- 1 year later …
- 1 of them did not actually have cancer ! (FP)
- Precision = 2/(2+1) = 67%
- 3 from the 97 healthy declared ones have cancer (FN)
- Recall = 2/(2+3) = 40%
- Accuracy = (94+2)/100 = 96%

Precision and Recall – examples

- A system which needs to launch a missile at a terrorist hideout located in a dense urban area.
- Precision not 100% → civilian casualties

- A system which needs to identify cancer-risk patients
- Recall not 100% → some patients will die of cancer

Precision and Recall – a probabilistic perspective

- n = # of patients who underwent a new cancer screening test
- Recall = Probability of test result + given a patient actually has cancer

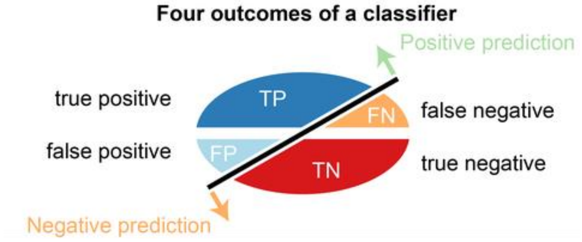
$$\frac{TP}{TP + FN}$$

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

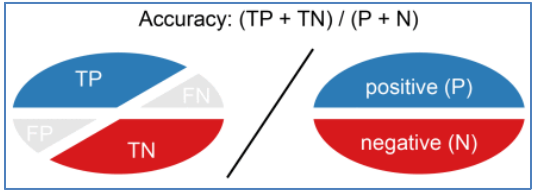
- Precision = Probability of actually having cancer given the test result is +

$$\frac{TP}{TP + FP}$$

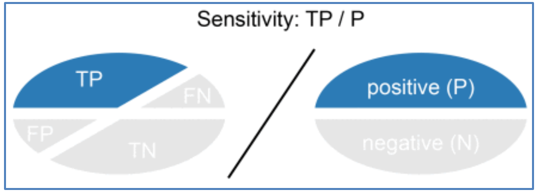
Summary of Measures



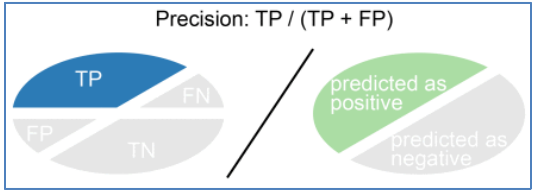
n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	



% of correct predictions



% of + class correctly predicted [aka Recall / TPR]



correct prediction of + class

% of – class incorrectly predicted

F1-score: A unified measure

- What to do when one classifier has better Precision but worse Recall, while other classifier behaves exactly opposite?
    - F-measure (Information Retrieval)
        - $F_1 = \dfrac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$

Utility and Cost

- What to do when one classifier has better Precision but worse Recall, while other classifier behaves exactly opposite?
  - F-measure (Information Retrieval)
    - $F_1 = \dfrac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$

→ **F1 measure punishes extreme values more !**

→ **Definition of Recall and Precision have same numerator, different denominators. A sensible way to combine them is harmonic mean.**

Utility and Cost

- Sometimes, there is a cost for each error
  - E.g. Earthquake prediction
    - False positive: Cost of preventive measures
    - False negative: Cost of recovery

- Detection Cost (Event detection)
  - Cost = C_{FP} * FP + C_{FN} * FN

x → Supervised Learning → y

Classification

Binary

{0,1}

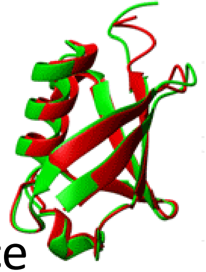
Multi-class

1-of-K

Multi-label

n-of-K

Structure



E.g. graph/sequence

Multi-class problems - Confusion matrix

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Avg. accuracy may not be very meaningful with imbalanced class label distribution

activity recognition from video

actual class

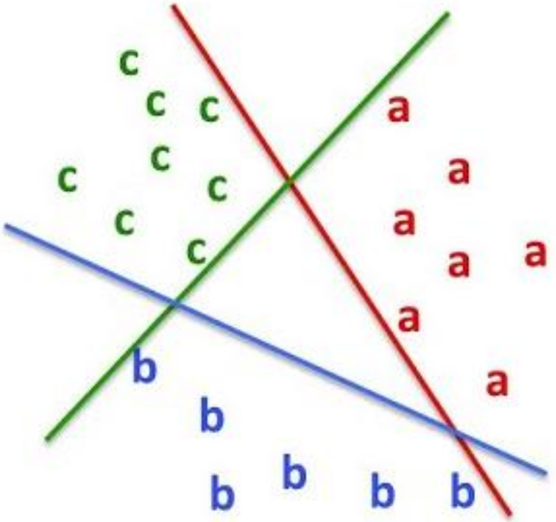
	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	100	0	0	0	0	0	0	0	0	0
jack	0	100	0	0	0	0	0	0	0	0
jump	0	0	89	0	0	0	11	0	0	0
pjump	0	0	0	100	0	0	0	0	0	0
run	0	0	0	0	89	0	11	0	0	0
side	0	0	0	0	0	100	0	0	0	0
skip	0	0	0	0	0	0	100	0	0	0
walk	0	0	0	0	0	0	0	100	0	0
wave1	0	0	0	0	0	0	0	0	67	33
wave2	0	0	0	0	0	0	0	0	0	100

predicted class

Courtesy: vision.jhu.edu

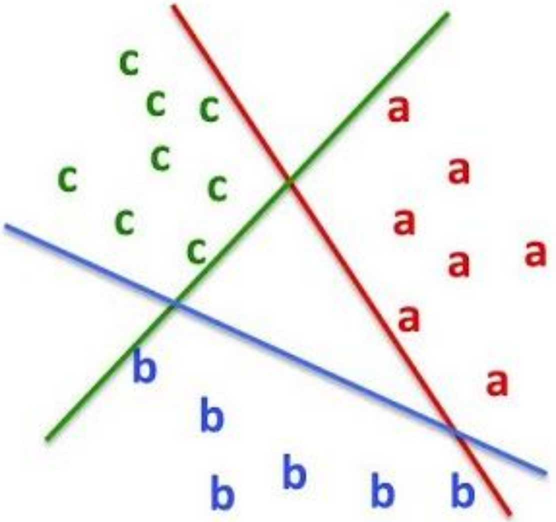
How to use 2-class measures for multi-class ?

- The `Cow-Essay’ strategy
  - Convert into 2-class problem(s) !

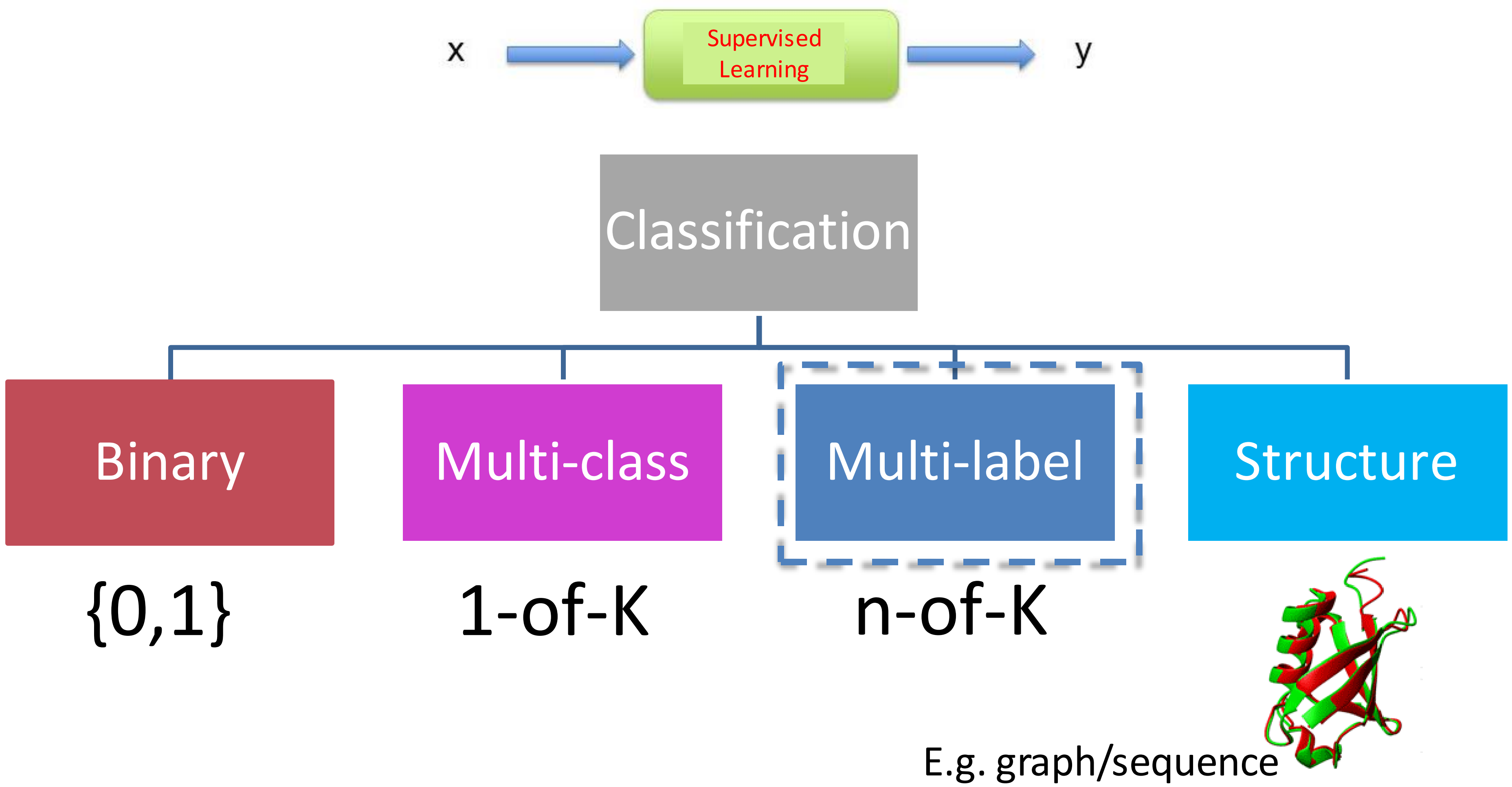


How to use 2-class measures for multi-class ?

- The `Cow-Essay' strategy
  - Convert into 2-class problem(s) !



- Average Precision, Recall etc.


x
Supervised Learning
y
Classification
Binary
Multi-class
Multi-label
Structure
{0,1}
1-of-K
n-of-K
E.g. graph/sequence

Example-based

- n is the number of examples.
- Y_i is the ground truth label assignment of the i^{th} example..
- x_i is the i^{th} example.
- $h(x_i)$ is the predicted labels for the i^{th} example.

Precision = $\frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap h(x_i)|}{|h(x_i)|}$

What fraction of labels are predicted correctly ?

Recall = $\frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap h(x_i)|}{|Y_i|}$

What % of correct labels were predicted ?

Accuracy = Fraction of samples predicted correctly

http://ieeexplore.ieee.org/document/6471714/

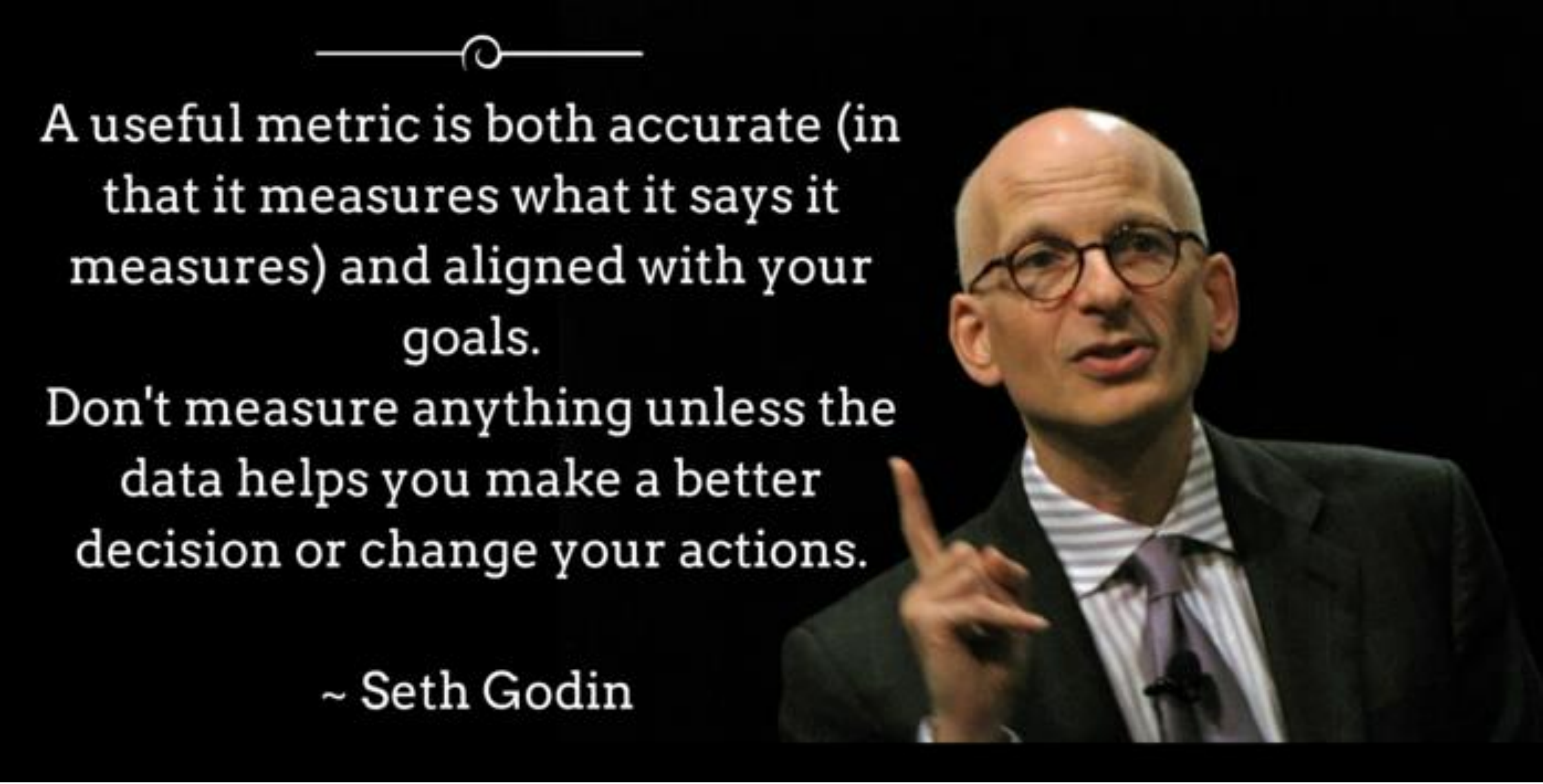
Summary

- Many metrics:
  - Accuracy, TP, FP, Precision, Recall, AP/mAP
  - Class imbalance and decision-cost imbalance must be taken into account

- Confusion Matrix: Important to analyze and refine solution.

Baselines

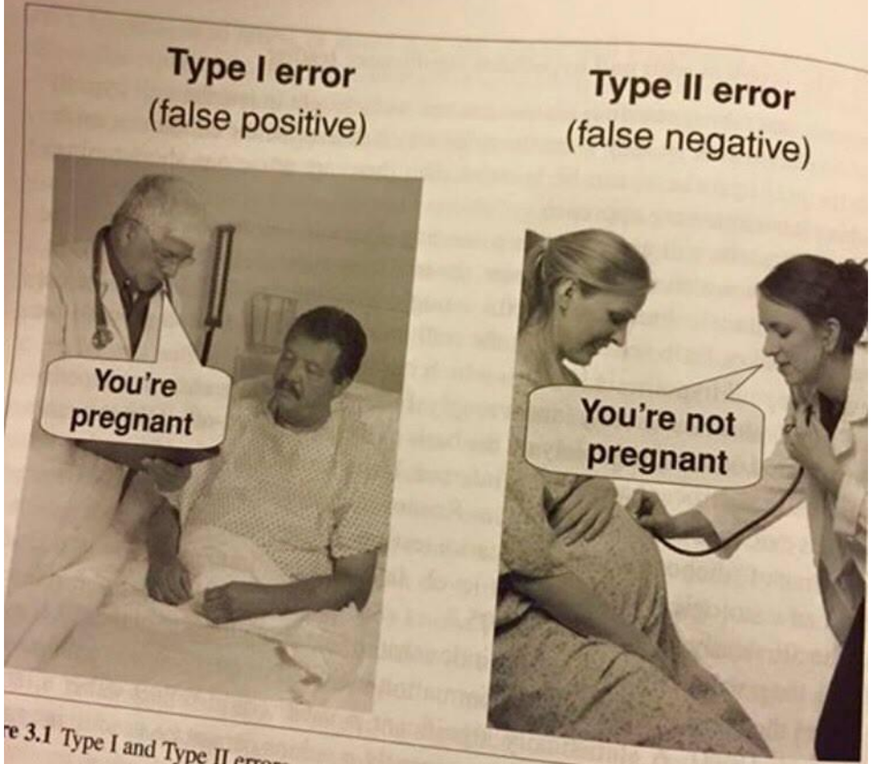
- 0 cost-to-build classifiers
- Binary
  - Equal # of samples / class → Random Guessing (50% accuracy)
  - Class imbalance
    - → Guess according to class proportion (Accuracy =)
    - 0-Rule: Majority class (Accuracy =) [slightly stronger baseline]


A useful metric is both accurate (in that it measures what it says it measures) and aligned with your goals.
Don't measure anything unless the data helps you make a better decision or change your actions.
~ Seth Godin

References and Reading

- https://classeval.wordpress.com/introduction/basic-evaluation-measures/
- https://towardsdatascience.com/what-metrics-should-we-use-on-imbalanced-data-set-precision-recall-roc-e2e79252aeba

- Code
  - https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics



re 3.1 Type I and Type II errors

levels to .01 or even .001