EVALUATING MACHINE LEARNING CLASSIFIERS: ACCURACY, PRECISION AND RECALL

Scott O'Hara
Metrowest Developers Machine Learning Group
07/11/2018

REFERENCES

Applied Machine Learning in Python

University of Michigan, Prof. Kevin Collins Thompson **(AMLP)**

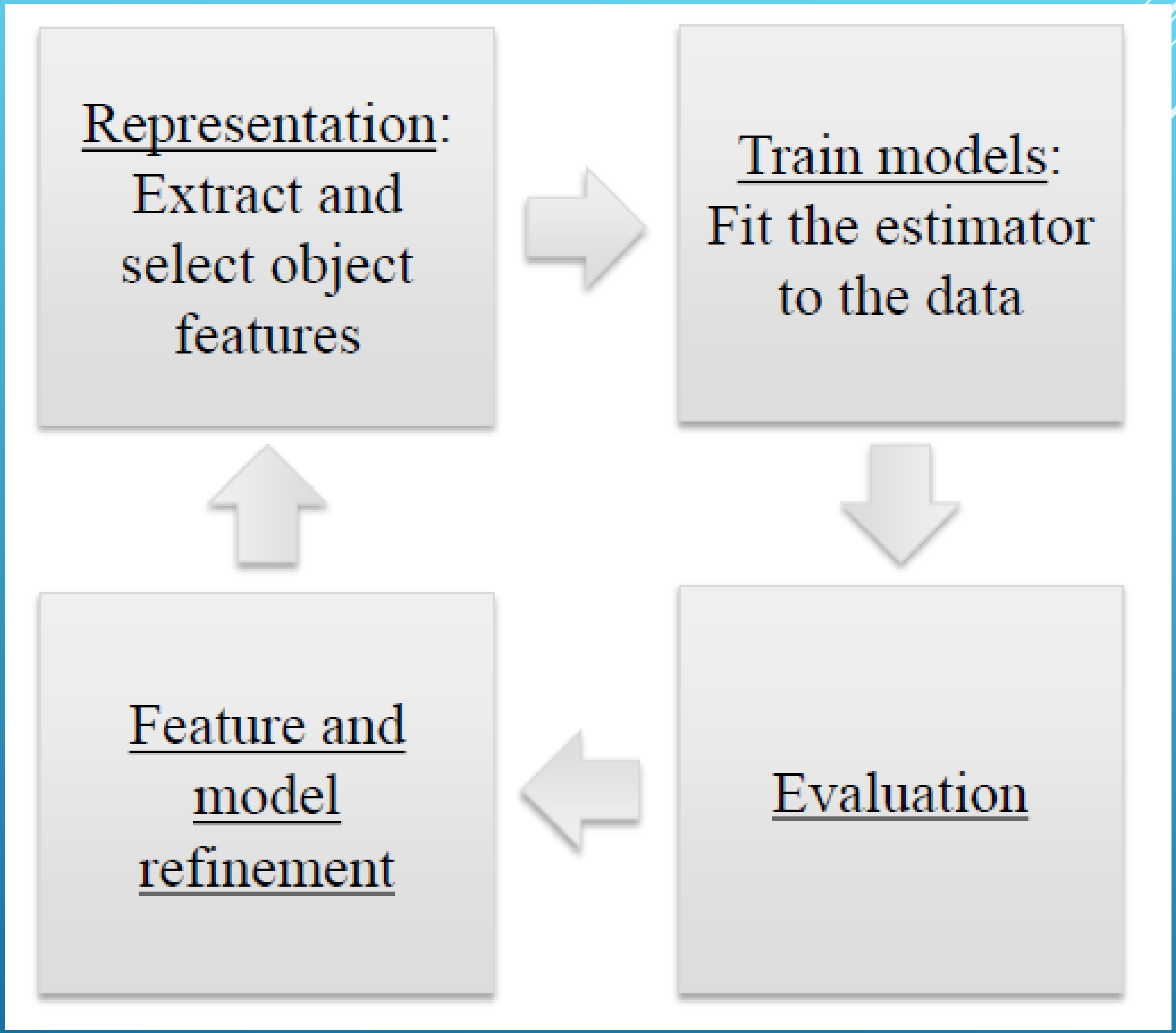
https://www.coursera.org/learn/python-machine-learning/home/welcome

Machine Learning: Classification

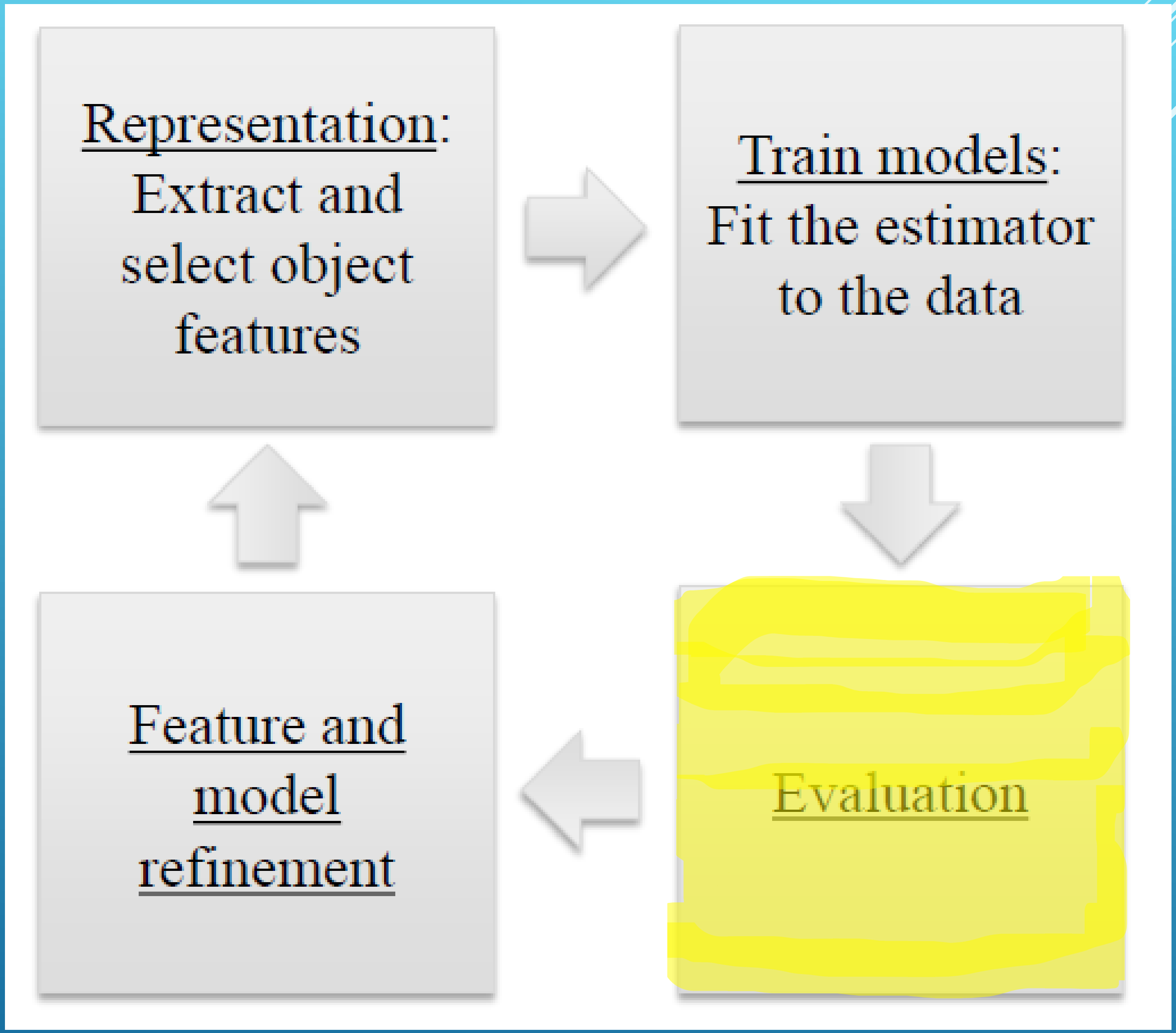
University of Washington, Profs. Emily Fox & Carlos Guestrin **(MLC)**

https://www.coursera.org/learn/ml-regression/home/welcome

REPRESENT, TRAIN, EVALUATE, REFINE



WHAT CRITERIA SHOULD WE USE TO EVALUATE OUR MODELS?



ACCURACY IS A COMMON METRIC

$$Accuracy = \frac{\#\ of\ correct\ predictions}{\#\ of\ total\ instances}$$

A model with 99.9% accuracy can sound really good!

HOWEVER, CONSIDER IMBALANCED CLASSES

- Suppose you have two classes:
  - Relevant (R): the positive class
  - Not_Relevant (N): the negative class
- Out of 1000 randomly selected items, on average
  - 1 item is relevant
  - 999 items are not relevant

A DUMMY CLASSIFIER GETS 99.9% ACCURACY!

- Classifier always predicts N
- Out of 1000 randomly selected items:

$$Accuracy = \frac{999}{1000} = 99.9\%$$

DUMMY CLASSIFIERS

- typically ignore training data features.
- often make predictions based on the distribution of the training data labels.
- can serve as a sanity check on your classifier's performance.

COMMON DUMMY CLASSIFIERS

most-frequent: predict most frequent label in training set.

stratified: random prediction based on training set distribution

uniform: choose predictions from a uniform distribution.

constant: predict constant label given by user.

EVALUATION

Different applications have different goals

Accuracy is widely used, but many other metrics are possible, e.g.,

- User satisfaction (Web search)

- Amount of revenue (e-commerce)

- Increase in patient survival rates (medical)

PRECISION AND RECALL

Two common alternatives to accuracy are: **precision** and **recall.**

PRECISION: fraction of positive predictions that are actually positive.

RECALL: fraction of positive examples that are predicted to be position

DOMAINS WHERE **PRECISION** IS IMPORTANT

- Search engine rankings, query suggestions
- Document classification
- Customer-facing tasks, e.g.,:
  - product recommendation
  - a restaurant website that automatically selects and posts positive reviews.

DOMAINS WHERE **RECALL** IS IMPORTANT

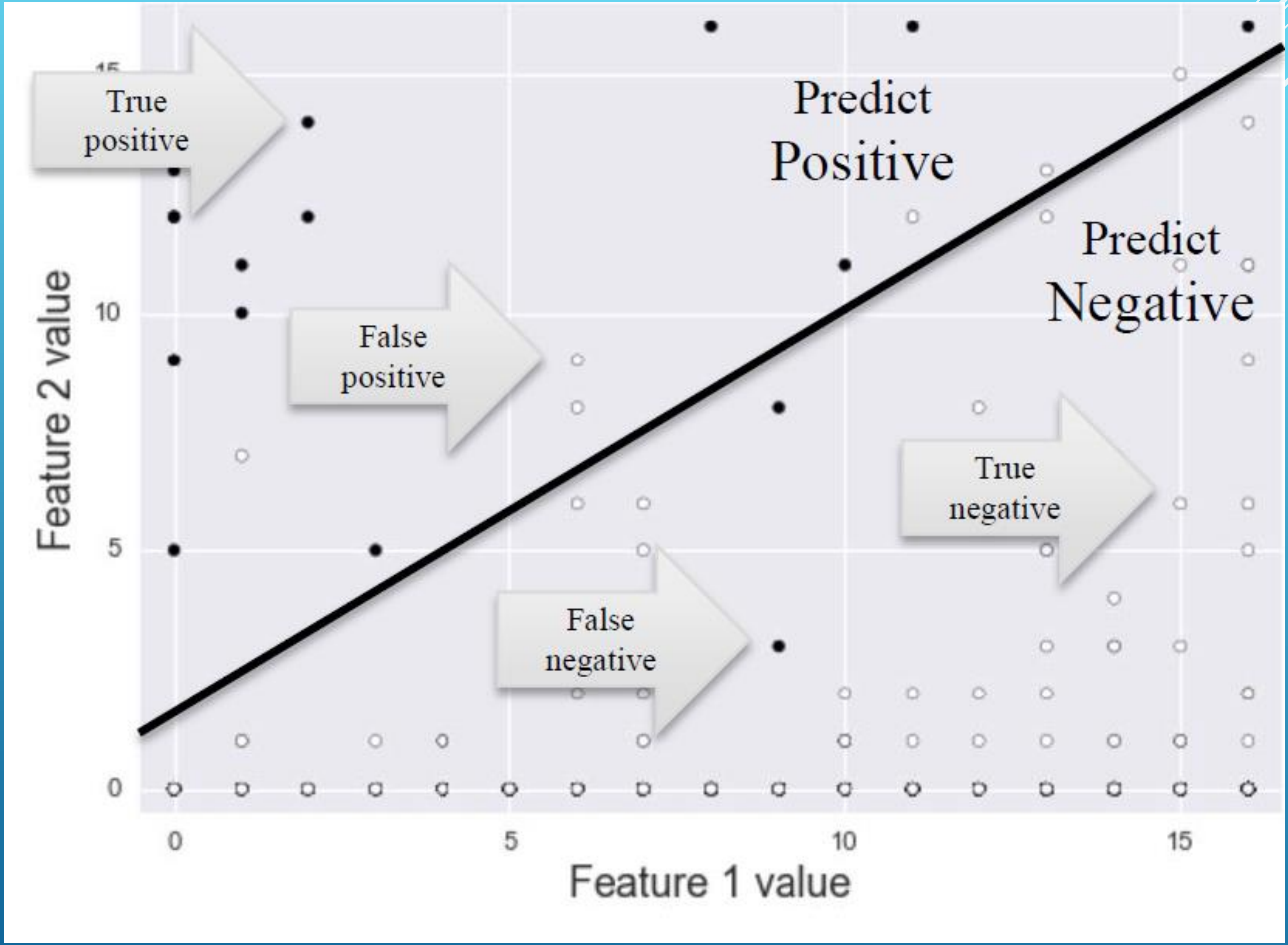
- Cancer tumor detection
- Search and information extraction in legal discovery.
- Often paired with a human expert to filter out false positives

	Predicted negative	**Predicted** positive
True negative	TN	FP
True positive	FN	TP

TP = true positive
FP = false positive
TN = true negative
FN = false negative

THE CONFUSION MATRIX

VISUALIZING DIFFERENT ERROR TYPES



THE CONFUSION MATRIX

	Predicted negative	Predicted positive
True negative	TN = 356	FP = 51
True positive	FN = 38	TP = 5

N = TN+TP+FN+FP = 450

- Every test instance is in exactly one box.
- Breaks down classifier results by error type (TP vs TN vs FP vs FN).
- Provides more information than simple accuracy.
- Helps you choose an evaluation metric that matches your project goals.
- There are many possible metrics that can be derived from the confusion matrix.

	Predicted negative	Predicted positive
True negative	TN	FP
True positive	FN	TP

- As FN + FP $\rightarrow$ 0, Accuracy $\rightarrow$ 1.0
- As FN + FP $\uparrow$, Accuracy $\rightarrow$ 0.0

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

ACCURACY

	Predicted negative	Predicted positive
True negative	TN	FP
True positive	FN	TP

- As FP → 0, Precision → 1.0
- As FP ↑, Precision → 0.0

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

PRECISION

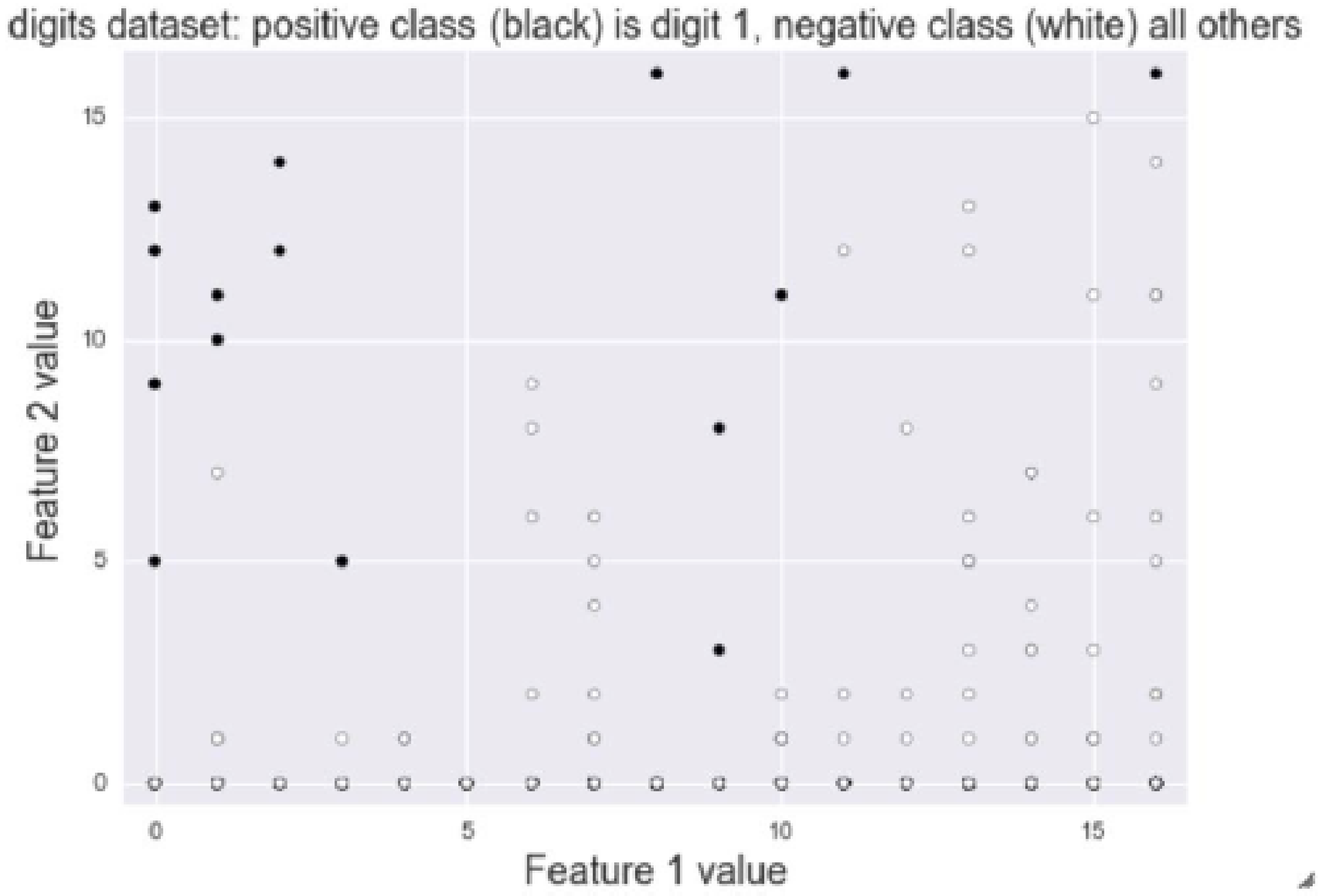
	Predicted negative	Predicted positive
True negative	TN	FP
True positive	FN	TP

- As FN → 0, Recall → 1.0
- As FN ↑, Recall → 0.0

$$\boldsymbol{Recall = \frac{TP}{TP + FN}}$$

RECALL

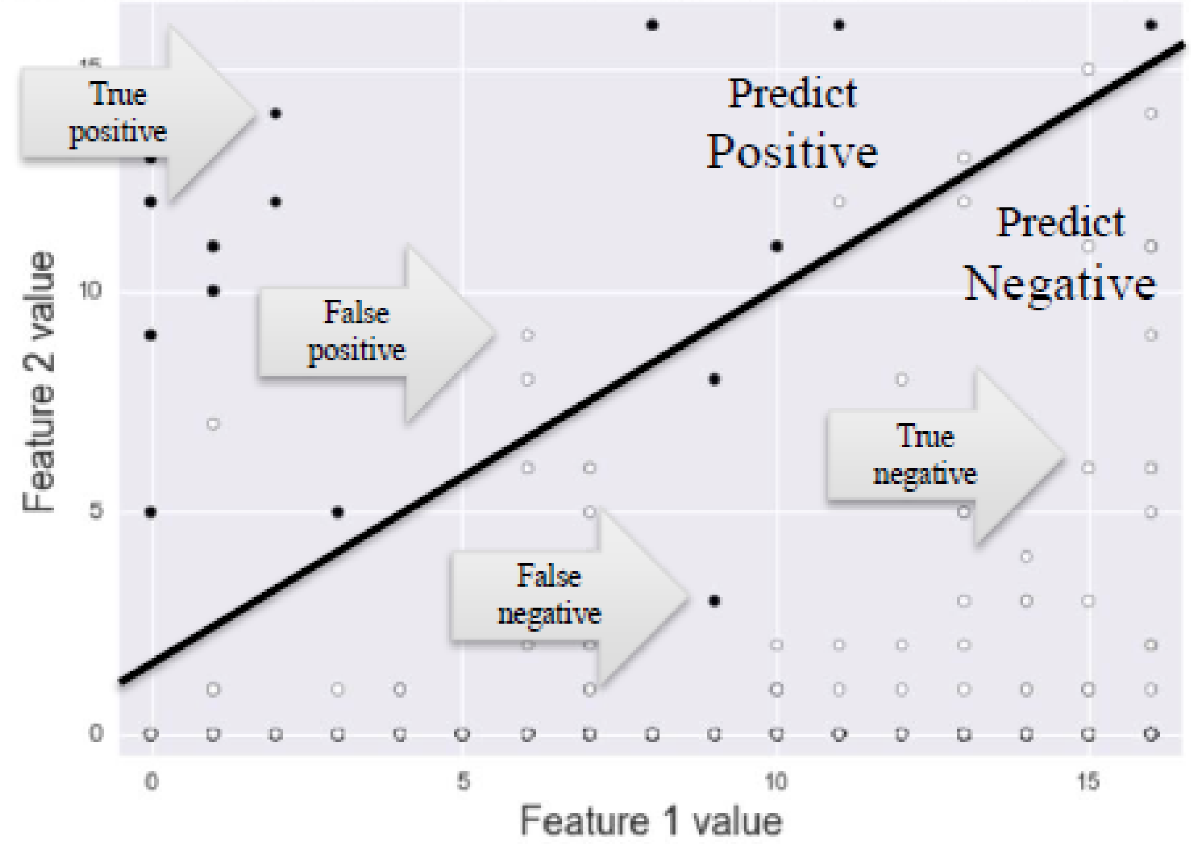
ILLUSTRATING PRECISION & RECALL



TN =	FP =
FN =	TP =

ILLUSTRATING PRECISION & RECALL

digits dataset: positive class (black) is digit 1, negative class (white) all others

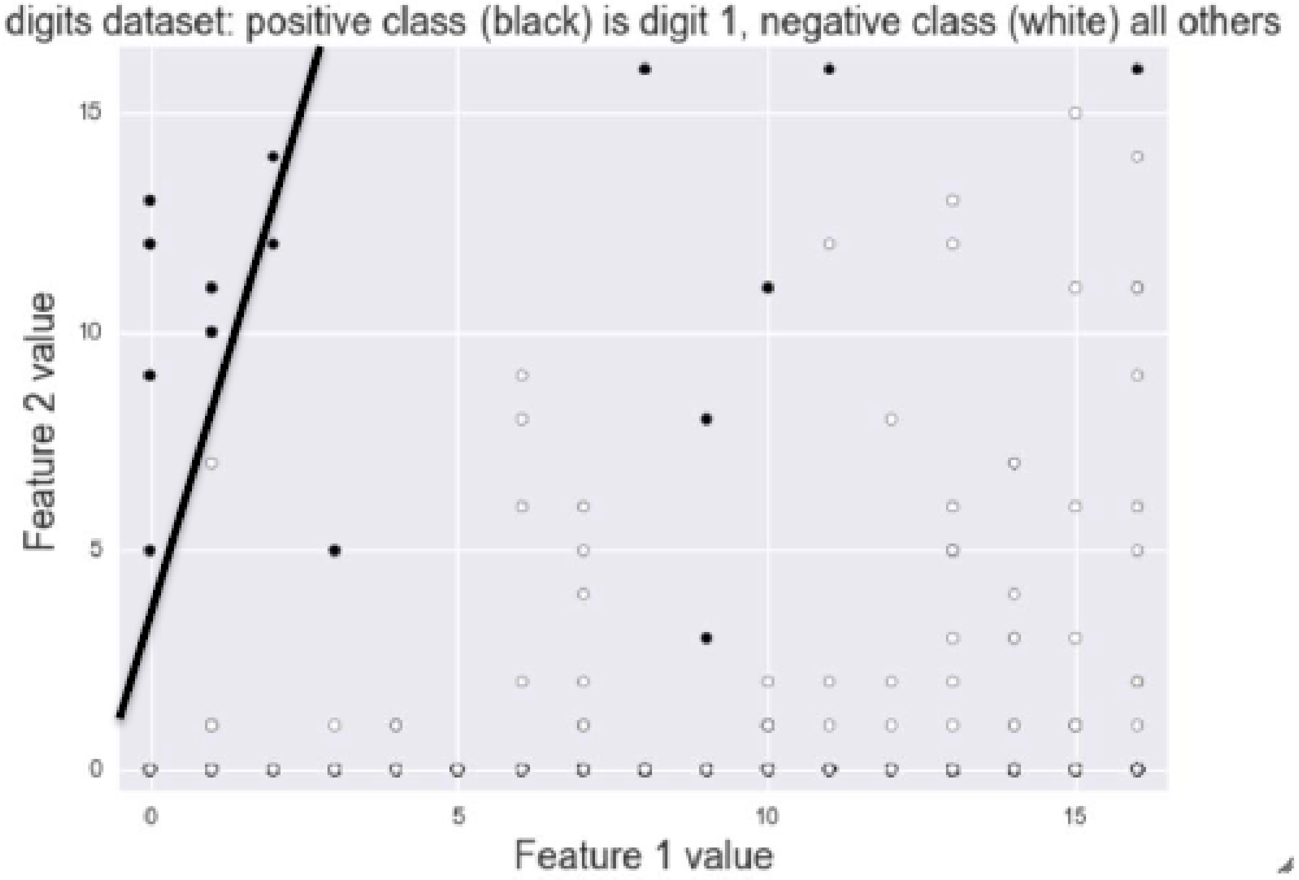


TN = 429	FP = 6
FN = 2	TP = 13

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{13}{19} = 0.68$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{13}{15} = 0.87$$

HIGH PRECISION / LOW RECALL

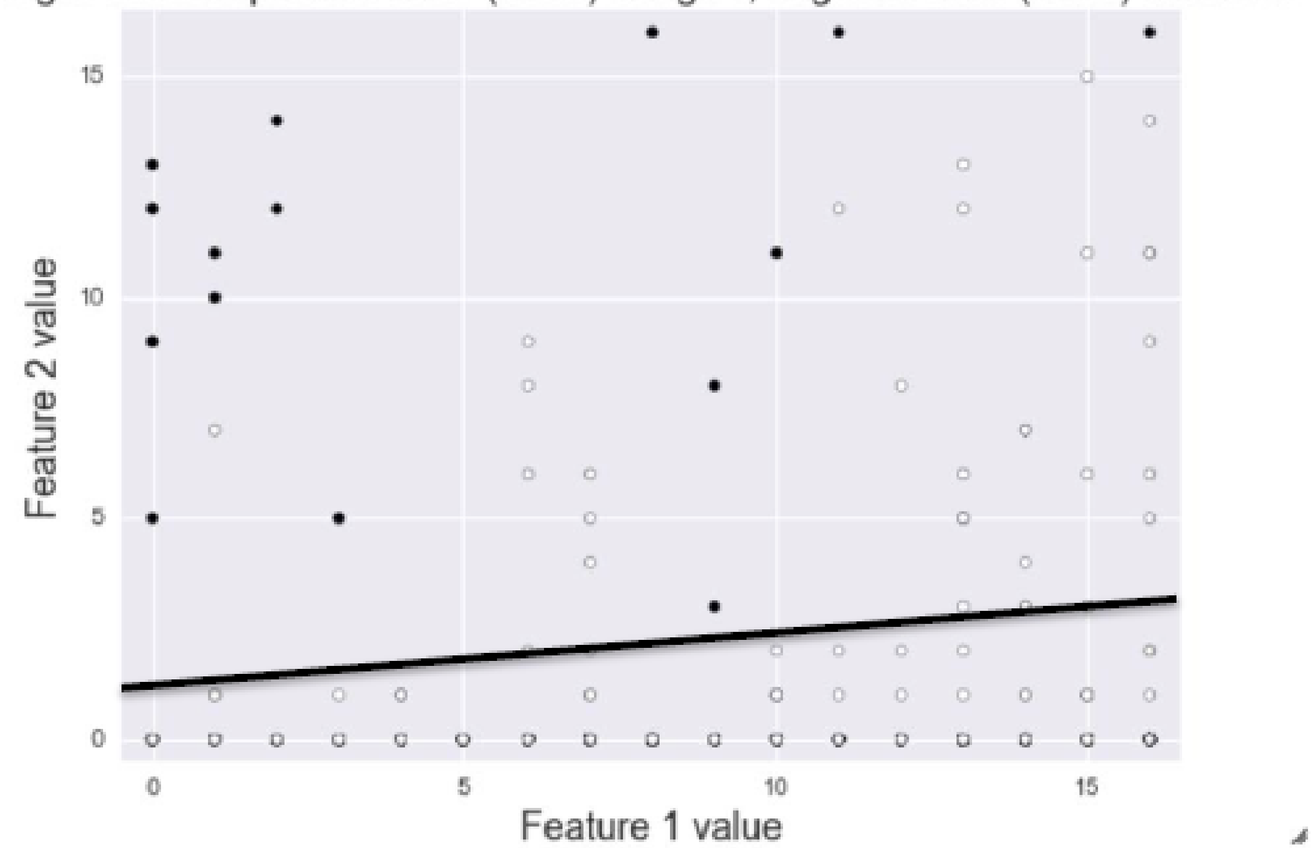


TN = 435	FP = 0
FN = 8	TP = 7

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{7}{7} = 1.00$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{7}{15} = 0.47$$

HIGH RECALL / LOW PRECISION



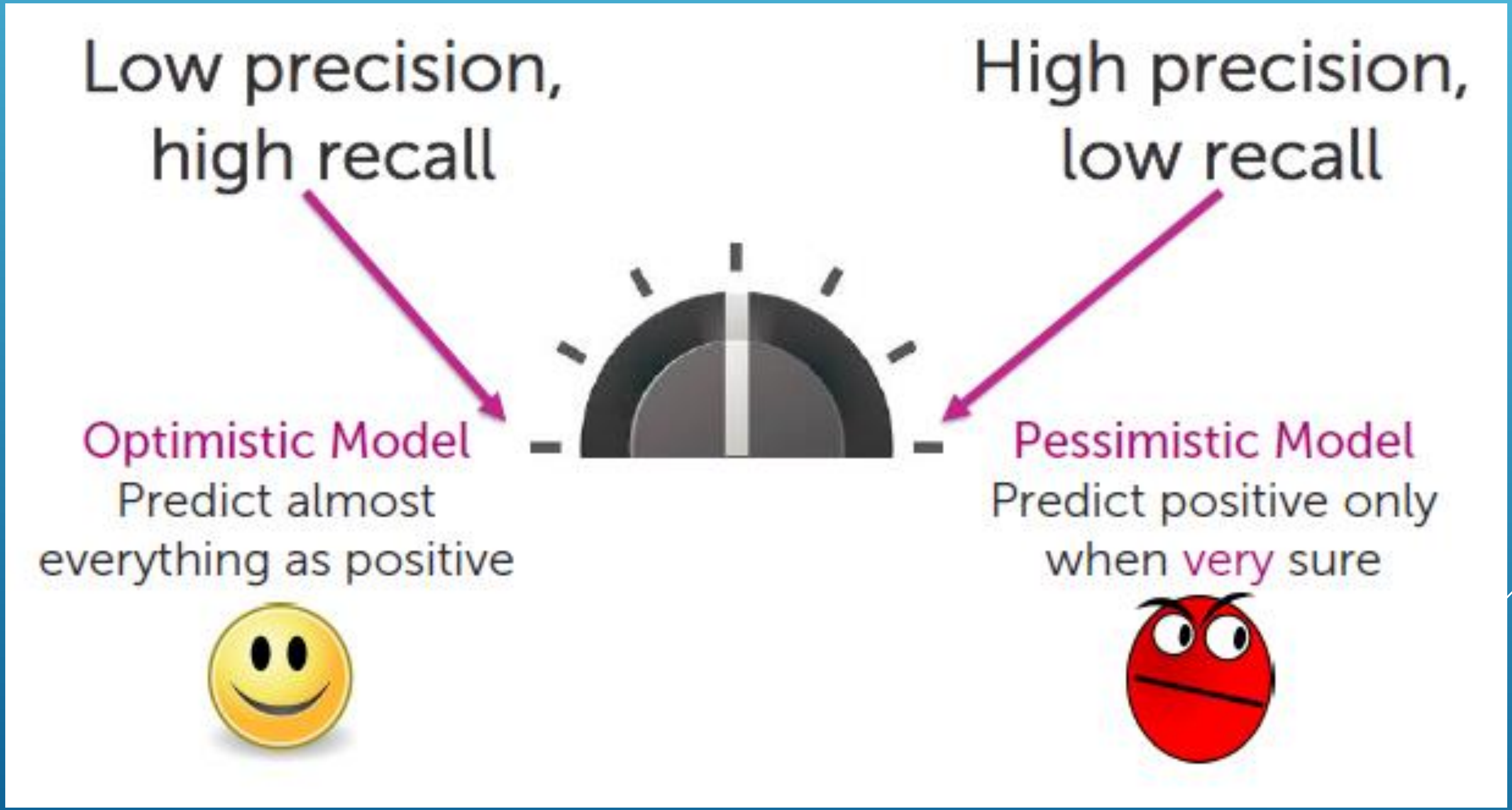
TN = 408	FP = 27
FN = 0	TP = 15

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{15}{42} = 0.36$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{15}{15} = 1.00$$

BALANCING PRECISION AND RECALL

Rather than seeking to maximize precision or recall, an optimal balance between the two is often sought.



THE F1-SCORE

- The **F1-score** combines precision and recall into a single number.
- The F1-score is the ***harmonic mean*** of precision and recall.

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FN + FP}$$

THE F-SCORE

- The **F-score** is a generalization of the F1-score.
- β allows adjustment of the metric to control the emphasis on recall vs precision.
  - β < 1.0 results in greater precision (minimize false positives)
  - β > 1.0 results in greater recall (minimize false negatives)

$$F_\beta = (1+\beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall} = \frac{(1+\beta^2) \cdot TP}{(1+\beta^2) \cdot TP + \beta \cdot FN + FP}$$

DECISION FUNCTIONS

- Any classifier that returns a score that represents how confident the classifier is in its prediction can be "adjusted" to result in a decision function that exhibits more or less precision or recall.

- By sweeping the decision threshold through the entire range of possible score values, we get a series of classification outcomes that form a curve.

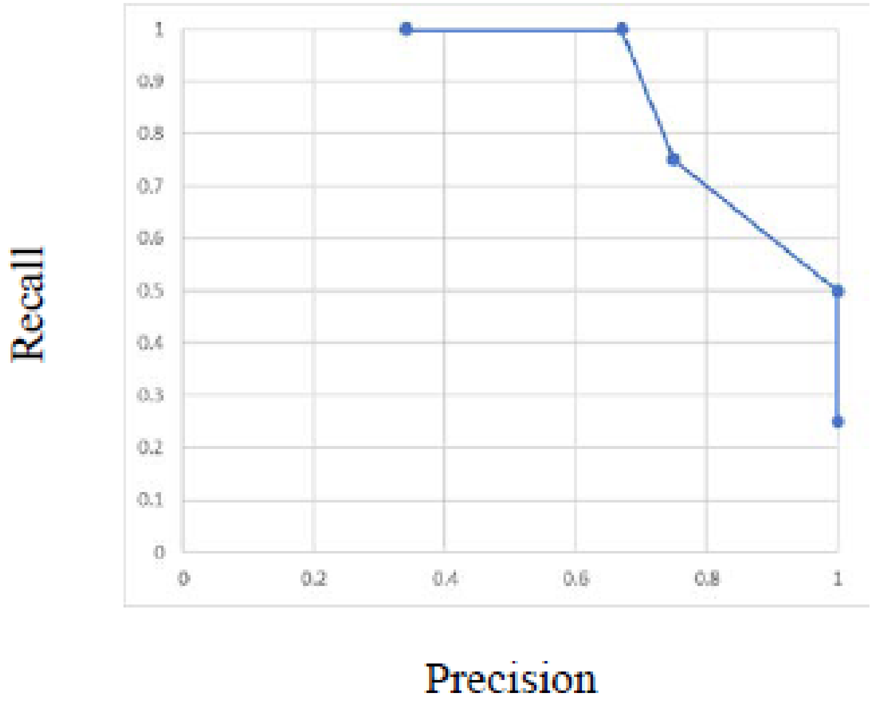
PROBABILISTIC CLASSIFIERS

- Some classifiers return a probability that an item is a particular class rather than a Boolean value.
- **Decision functions** can be constructed from probabilistic classifier.
- Examples include **Logistic regression**, **Naïve Bayes**.
- Typical rule is choose likely class if $P(x) >$ *threshold* where threshold = 0.5
- Adjusting *threshold* affects predictions of classifier
- A higher *threshold* results in a more "pessimistic" classifier i.e., it increase precision.
- A lower *threshold* results in a more "optimistic" classifier i.e., it increases recall.

VARYING THE DECISION THRESHOLD

True Label	Classifier score
0	-27.6457
0	-25.8486
0	-25.1011
0	-24.1511
0	-23.1765
0	-22.575
0	-21.8271
0	-21.7226
0	-19.7361
0	-19.5768
0	-19.3071
0	-18.9077
0	-13.5411
0	-12.8594
1	-3.9128
0	-1.9798
1	1.824
0	4.74931
1	15.234624
1	21.20597

Classifier score threshold	Precision	Recall
-20	4/12=0.34	4/4=1.00
-10	4/6=0.67	4/4=1.00
0	3/4=0.75	3/4=0.75
10	2/2=1.0	2/4=0.50
20	1/1=1.0	1/4 = 0.25

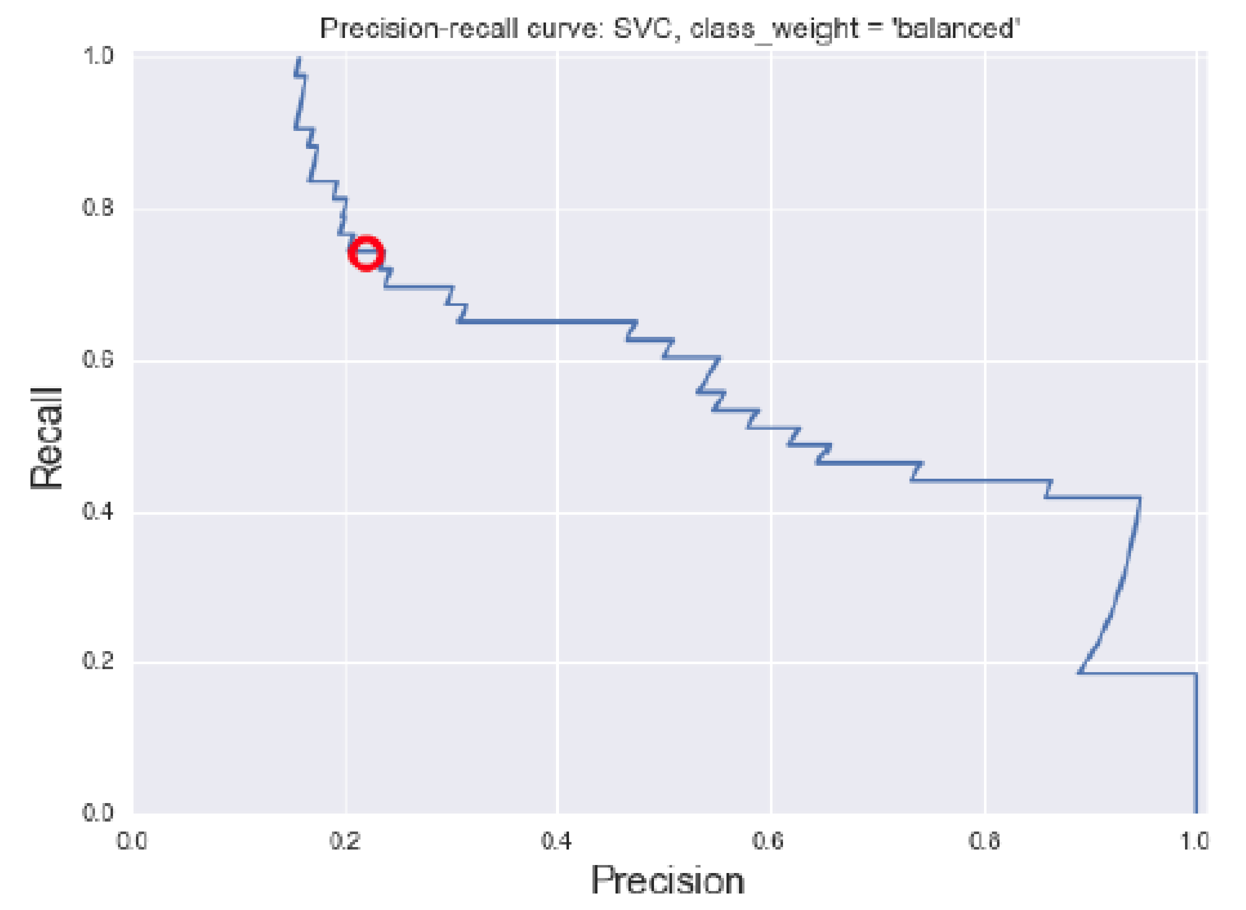


PRECISION-RECALL CURVES

X-axis: Precision
Y-axis: Recall

Top right corner:
- **The “ideal” point**
- **Precision = 1.0**
- **Recall = 1.0**

“Steepness” of P-R curves is important:
- **Maximize precision**
- **while maximizing recall**



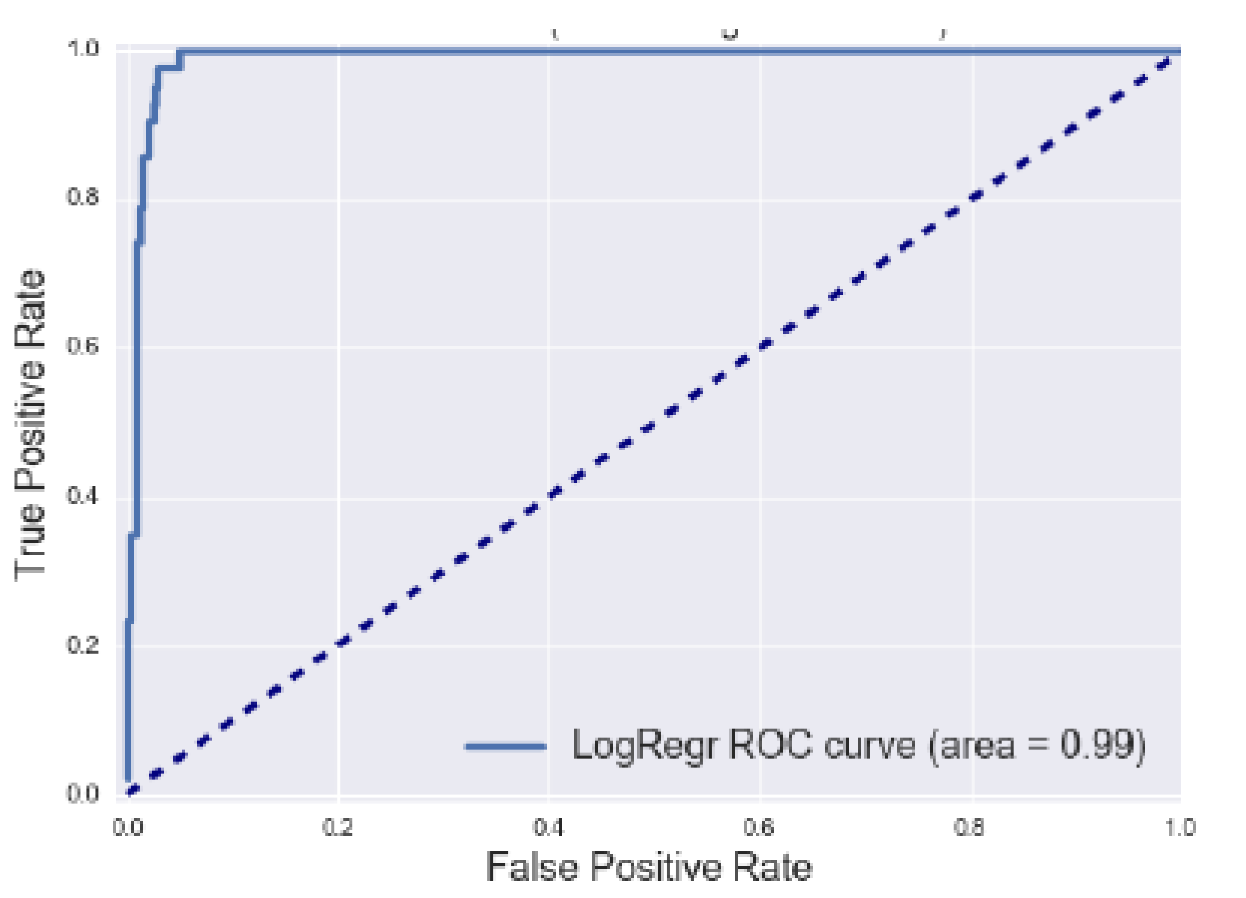
AMLP

ROC CURVES

X-axis: False Positive Rate
Y-axis: True Positive Rate

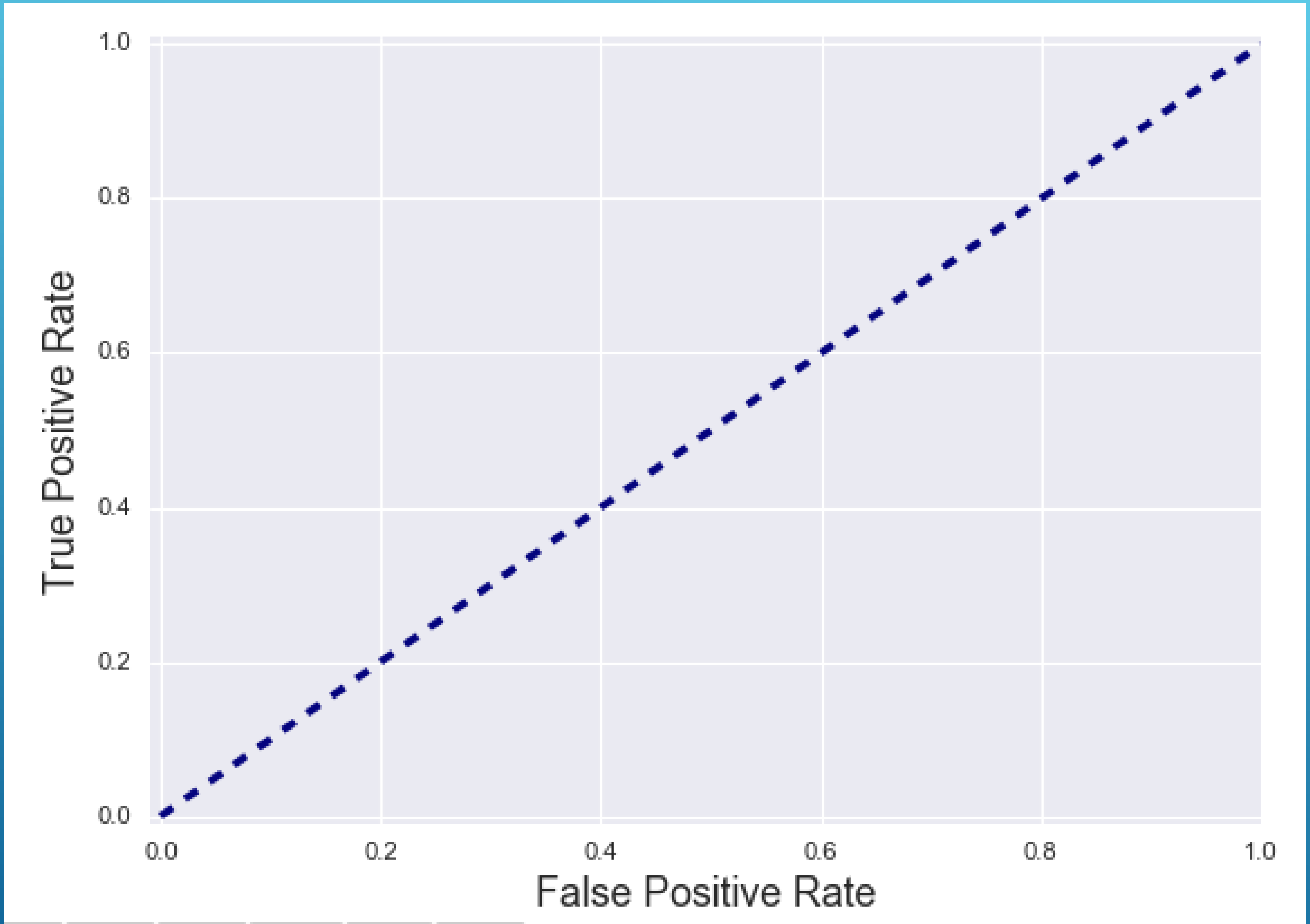
Top left corner:
- **The "ideal" point**
- **False positive rate of zero**
- **True positive rate of one**

"Steepness" of ROC curves is important:
- **Maximize the true positive rate**
- **while minimizing the false positive rate**

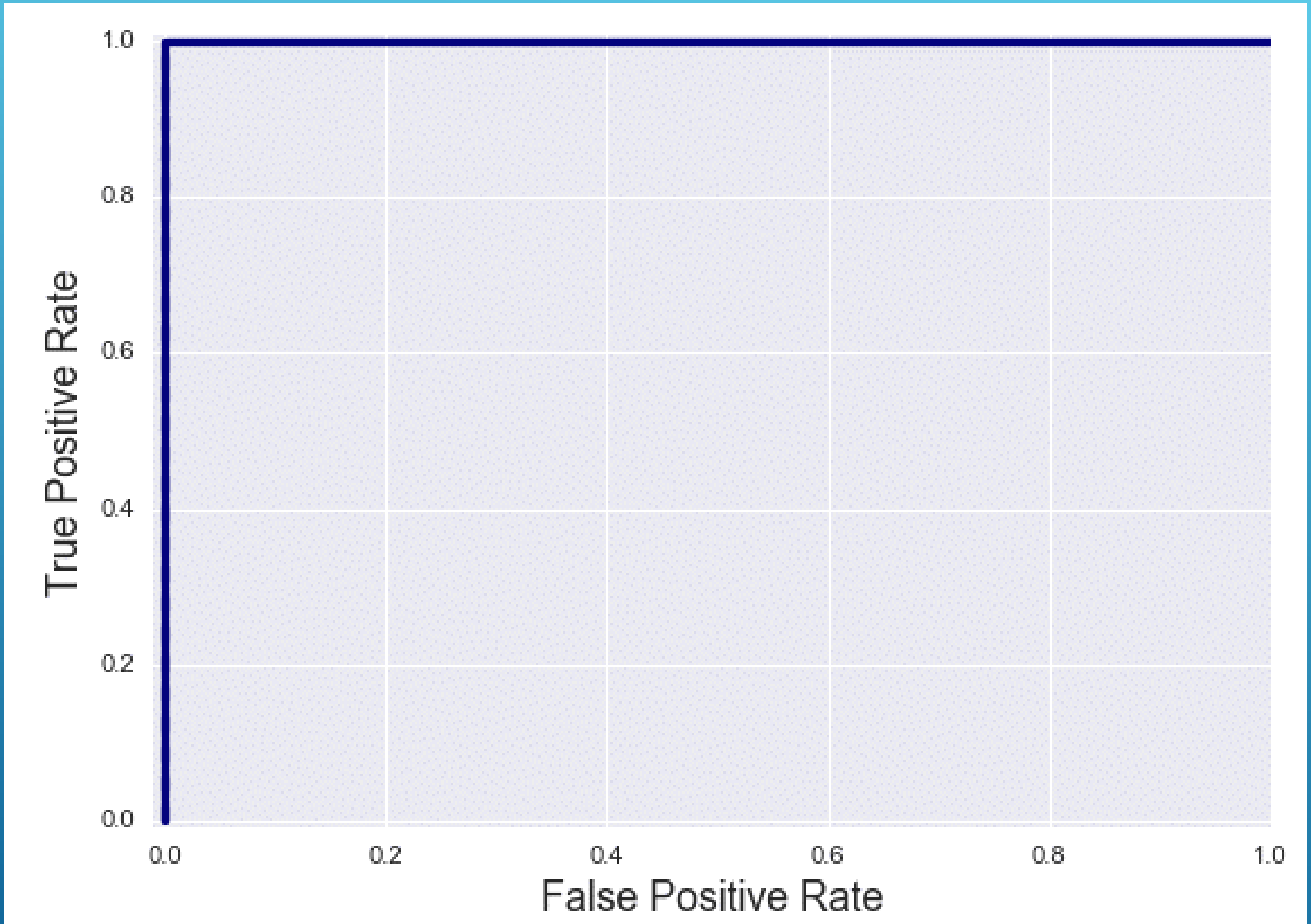


ROC = Receiver Operating Characteristic

ROC CURVES: RANDOM GUESSING

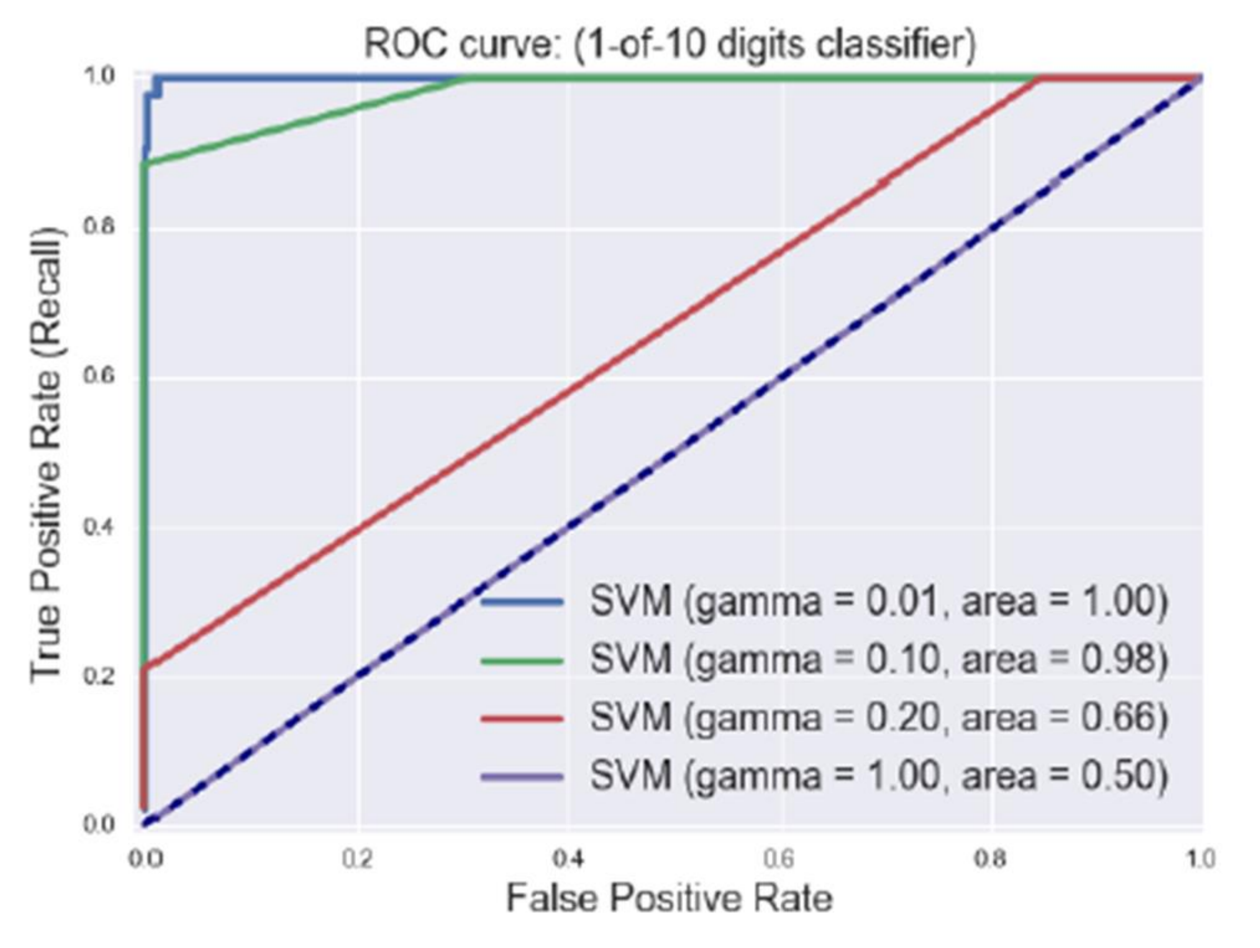


ROC CURVES: PERFECT CLASSIFIER



SUMMARIZING AN ROC CURVE IN ONE NUMBER: AREA UNDER THE CURVE (AUC)

- AUC = 0 (worst) AUC = 1 (best)
- AUC can be interpreted as:
  1. The total area under the ROC curve.
  2. The probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example.
- Advantages:
  – Gives a single number for easy comparison.
  – Does not require specifying a decision threshold.
- Drawbacks:
  – As with other single-number metrics, AUC loses information, e.g. about tradeoffs and the shape of the ROC curve.
  – This may be a factor to consider when e.g. wanting to compare the performance of classifiers with overlapping ROC curves.



AMLP

CONCLUSION

- Accuracy is often not the right evaluation metric for many real-world machine learning tasks
- False positives and false negatives sometimes need to be treated very differently
- Make sure you understand the needs of your application and choose an evaluation metric that matches your application, user, or business goals.

A GENERAL APPROACH

1. Consider carefully the data you have and what you are trying to do with it.
2. Choose a SINGLE metric and optimize that metric.
3. If this gives satisfactory results, then you are done. Otherwise return to step 1.

OTHER EVALUATION METHODS

- **Learning curve**: How much does accuracy (or other metric) change as a function of the amount of training data?
- **Sensitivity analysis**: How much does accuracy (or other metric) change as a function of key learning parameter values?