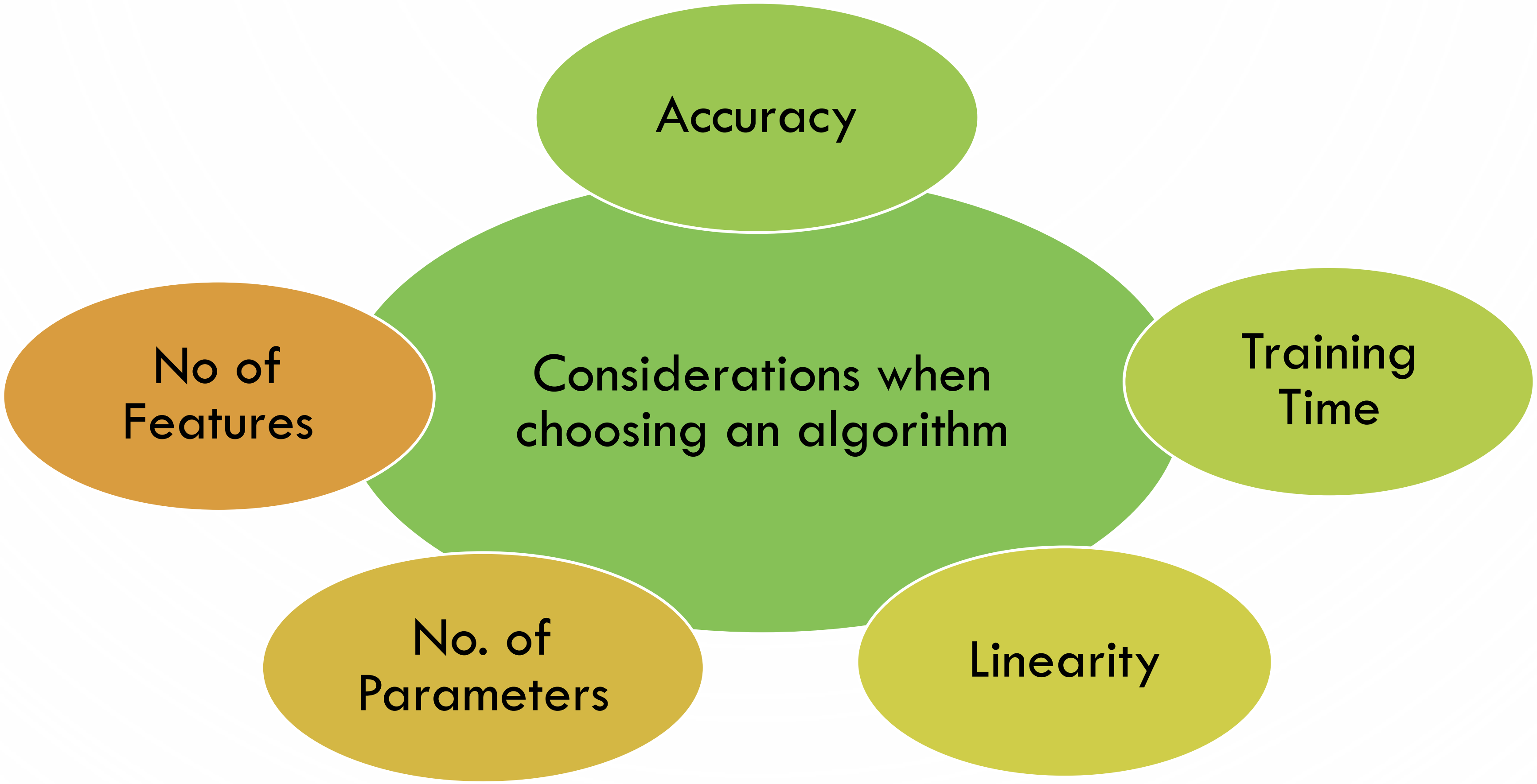
# MODEL EVALUATION AND SELECTION

Considerations when choosing a machine learning algorithm:

1. Accuracy
2. Training Time
3. Linearity
4. Number of Parameters
5. Number of features

# MODEL EVALUATION AND SELECTION



# MODEL EVALUATION AND SELECTION

- ACCURACY
  - IT IS NOT ALWAYS NECESSARY TO GET ACCURATE RESULTS
  - APPROXIMATION IS SOMETIMES SUFFICIENT
  - IT CUTS PROCESSING TIME SIGNIFICANTLY
  - TENDS TO AVOID OVERFITTING
- TRAINING TIME
  - TIME TO TRAIN A MODEL VARIES GREATLY ACROSS ALGORITHM
  - IT IS HIGHLY CORRELATED TO ACCURACY
  - DEPENDENT ON THE SIZE OF THE TRAINING DATA SET AS WELL

# MODEL EVALUATION AND SELECTION

- LINEARITY
  - LINEAR CLASSIFICATION ALGORITHM ASSUMES CLASSES CAN BE SEPARATED LINEARLY
  - HOWEVER, IF THE DATA ARE NOT LINEARLY SEPARABLE, IT MAY RESULT IN LOW ACCURACY
- NO OF PARAMETERS
  - NO OF PARAMETER DENOTES THE FLEXIBILITY OF AN ALGORITHM
  - WHEN THE RIGHT COMBINATION OF PARAMETERS IS YIELD, IT WILL BRING ABOUT HIGH ACCURACY.
  - HOWEVER, REQUIRES A LOT OF TRIAL AND ERROR WORK.
- NO OF FEATURES
  - IN SOME DATASET, NO OF FEATURES CAN BE VERY LARGE COMPARED TO THE NUMBER OF DATA POINTS
  - THIS CHARACTERISTIC MAY BOG DOWN SOME MACHINE LEARNING ALGORITHMS (RESULTING IN SUBSTANTIAL TRAINING TIME)
  - SVM USUALLY HANDLES THESE KIND OF DATA WELL.

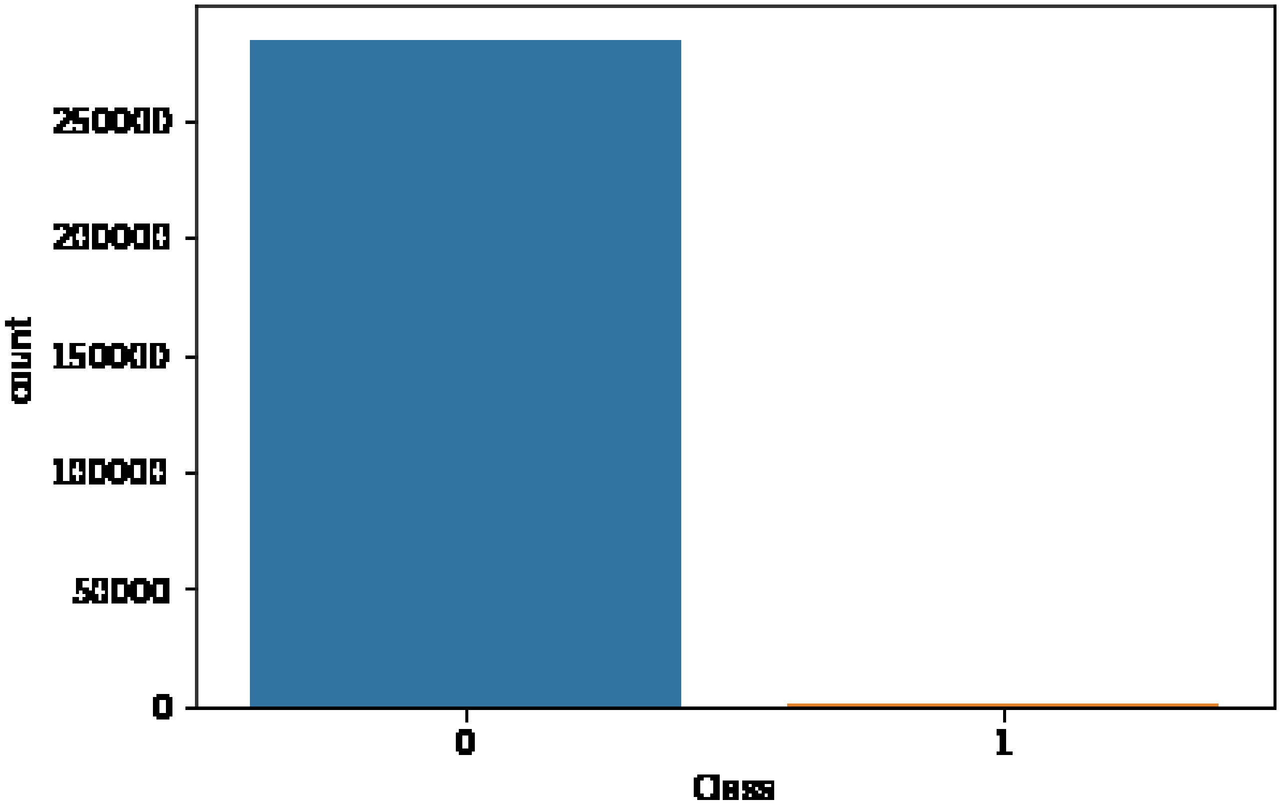
# EVALUATION

- IT'S VERY IMPORTANT TO CHOOSE EVALUATION METHODS THAT MATCH THE GOAL OF YOUR APPLICATION.
- COMPUTE YOUR SELECTED EVALUATION METRIC FOR MULTIPLE DIFFERENT MODELS.
- THEN SELECT THE MODEL WITH 'BEST' VALUE OF EVALUATION METRIC.

# EVALUATION

ACCURACY = #CORRECT PREDICTIONS /#TOTAL INSTANCES

# ACCURACY WITH IMBALANCED CLASSES



# CONFUSION MATRIX

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

Label 1 = positive class (class of interest)

Label 0 = negative class (everything else)

TP = true positive
FP = false positive (Type I error)
TN = true negative
FN = false negative (Type II error)

# CONFUSION MATRIX

| | | | |
|---|---|---|---|
| True negative | TN = 400 | FP = 7 | |
| True positive | FN = 17 | TP = 26 | |
| | Predicted negative | Predicted positive | $N = 450$ |

# CONFUSION MATRIX

| | | | |
|---|---|---|---|
| True negative | TN = 400 | FP = 7 | |
| True positive | FN = 17 | TP = 26 | |
| | Predicted negative | Predicted positive | $N = 450$ |

# ACCURACY

Accuracy: for what fraction of all instances is the classifier's prediction correct (for either positive or negative class)?

| | | | |
|---|---|---|---|
| True negative | TN = 400 | FP = 7 | |
| True positive | FN = 17 | TP = 26 | |
| | Predicted negative | Predicted positive | $N = 450$ |

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

$$= \frac{400+26}{400+26+17+7}$$

$$= 0.95$$

# RECALL

## Recall or True Positive Rate (TPR)

Recall, or True Positive Rate (TPR): what fraction of all positive instances does the classifier <u>correctly</u> identify as positive?

| | | | |
|---|---|---|---|
| True negative | TN = 400 | FP = 7 | |
| True positive | FN = 17 | TP = 26 | |
| | Predicted negative | Predicted positive | $N = 450$ |

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

$$= \frac{26}{26+17}$$

$$= 0.60$$

Recall is also known as:

- True Positive Rate (TPR)
- Sensitivity
- Probability of detection

# PRECISION

Precision: what fraction of positive predictions are correct?

| | | | |
|---|---|---|---|
| True negative | TN = 400 | FP = 7 | |
| True positive | FN = 17 | TP = 26 | |
| | Predicted negative | Predicted positive | $N = 450$ |

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

$$= \frac{26}{26+7}$$

$$= 0.79$$

# SPECIFICITY

## Specificity or False Positive Rate (FPR)

False positive rate (FPR): what fraction of all negative instances does the classifier incorrectly identify as positive?

| | Predicted negative | Predicted positive | |
|---|---|---|---|
| True negative | TN = 400 | FP = 7 | |
| True positive | FN = 17 | TP = 26 | |
| | | | $N = 450$ |

$$FPR = \frac{FP}{TN+FP}$$

$$= \frac{7}{400+7}$$

$$= 0.02$$

False Positive Rate is also known as:

- Specificity

# PRECISION VS RECALL

- Recall-oriented machine learning tasks:
- ❑ • Search and information extraction in legal discovery
- ❑ • Tumor detection
- ❑ • Often paired with a human expert to filter out false positives
- ❑ • Precision-oriented machine learning tasks:
- ❑ • Search engine ranking, query suggestion
- ❑ • Document classification
- ❑ • Many customer-facing tasks (users remember failures!)

# F1-SCORE

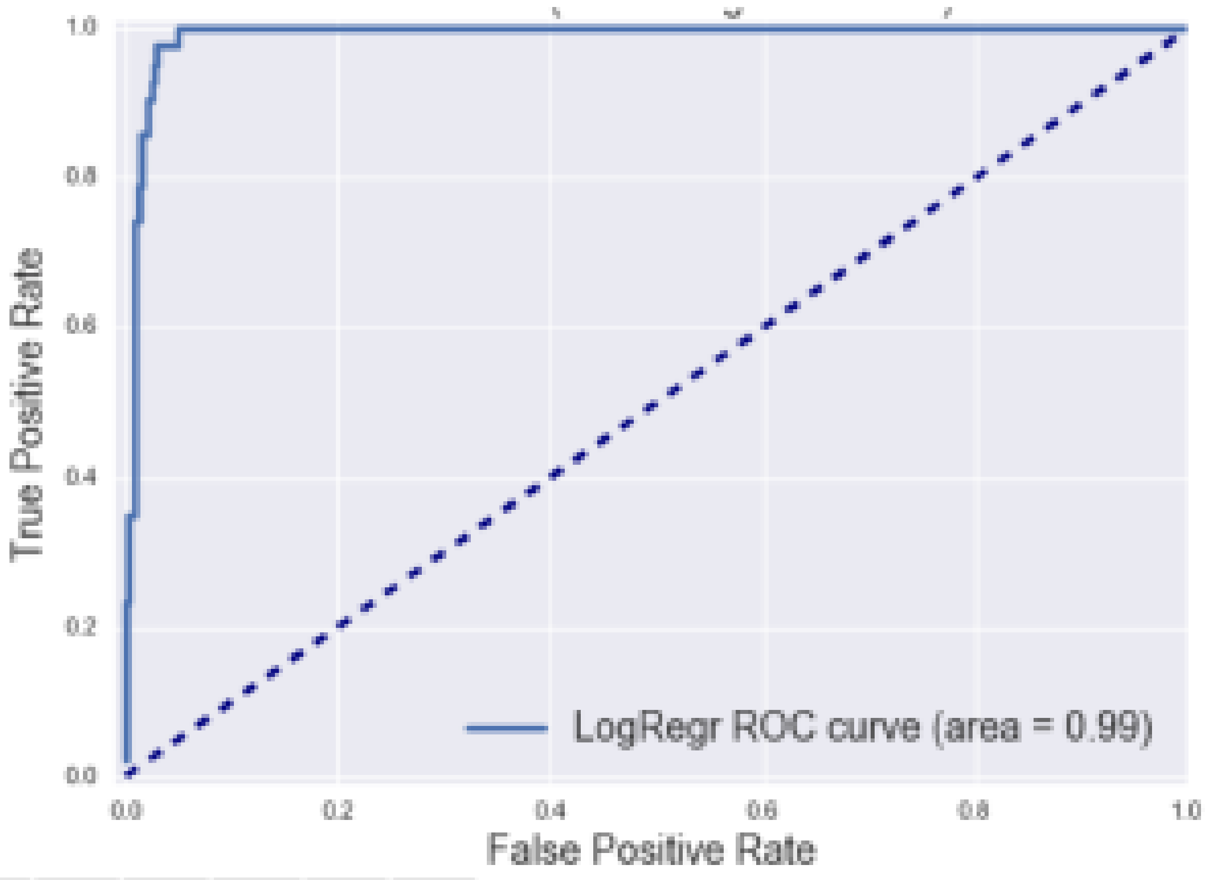
$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FN + FP}$$

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

$\beta$ allows adjustment of the metric to control the emphasis on recall vs precision:

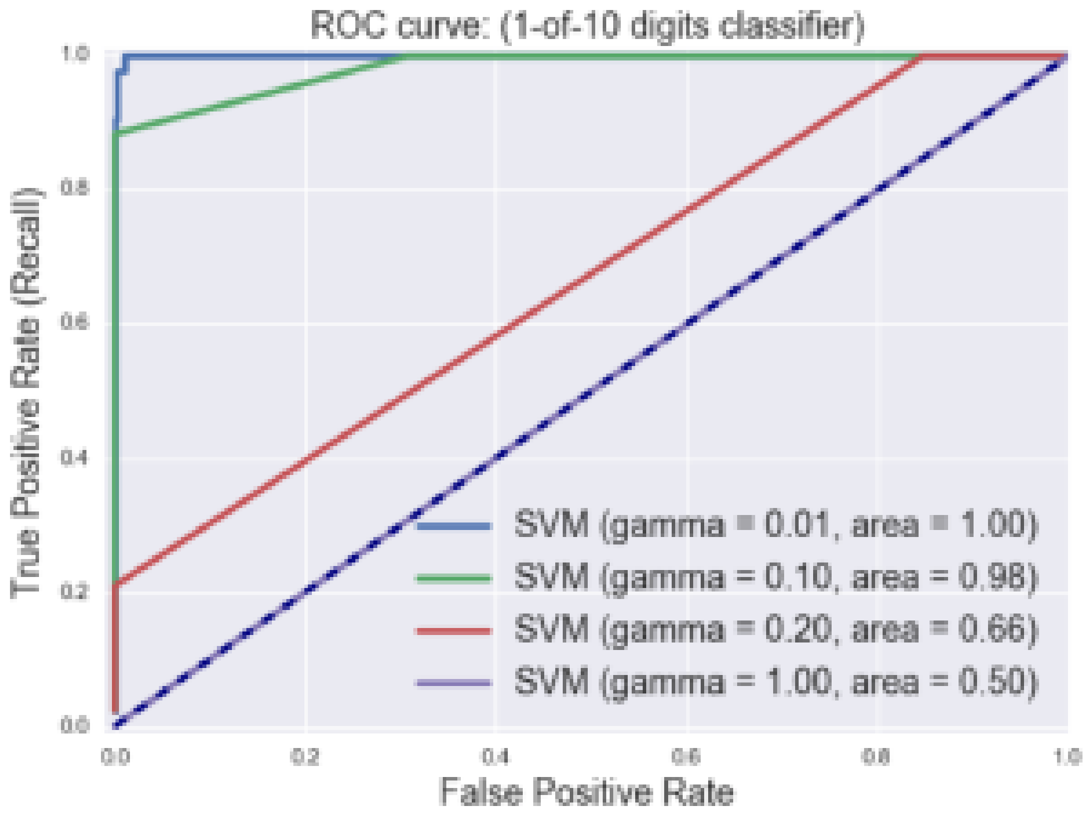
- Precision-oriented users: = 0.5 (false positives hurt performance more than false negatives)
- Recall-oriented users: = 2 (false negatives hurt performance more than false positives)

# ROC CURVE




- 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

# 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.

# MODEL SELECTION

**Train/Test on same data**

- Single Metric
- Typically overfits and likely won't generalize well to new data
- But can serve as sanity check: low accuracy on the training set may indicate an implementation problem
- **Single train/test split**
- Single Metric
- Speed and simplicity
- Lack of variance information
- **K-fold cross validation**
- K train-test splits
- Average metric over all splits
- Can be combined with parameter grid search: GridSearchCV (default cv = 3)