AI/ML for Environmental Data Analytics

DS 392

Deepak Subramani
Assistant Professor
Dept. of Computational and Data Science
Indian Institute of Science Bengaluru

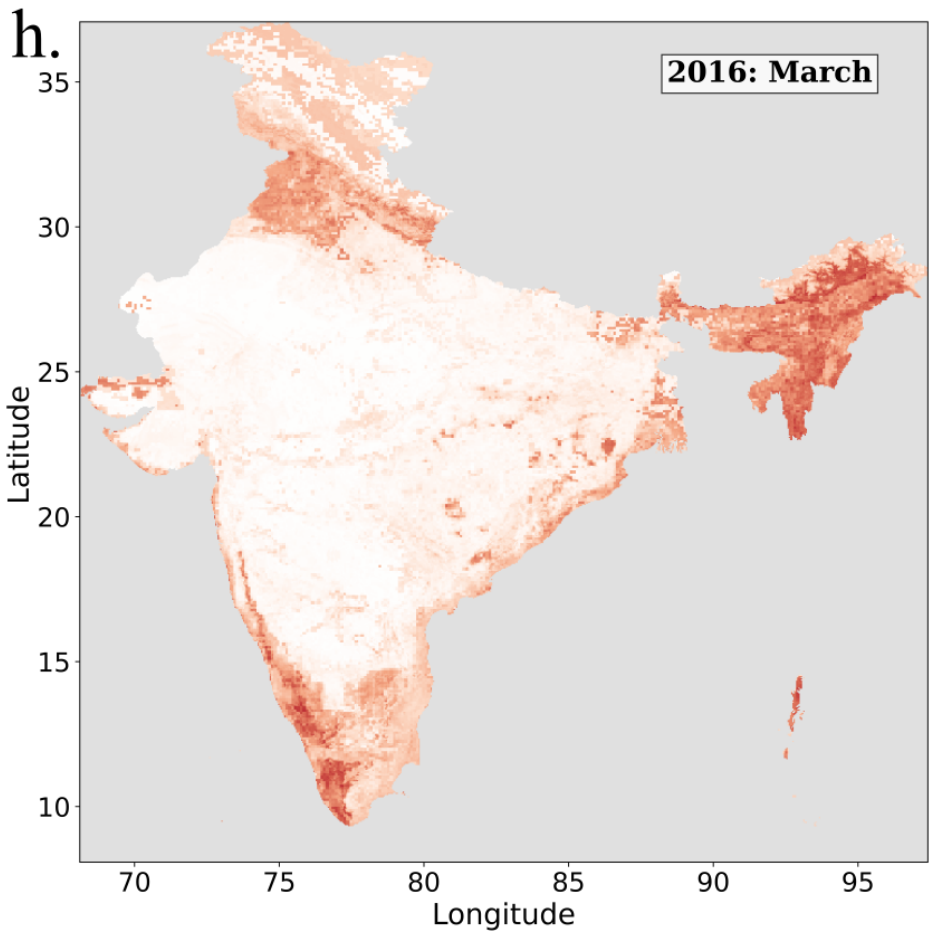
The Machine Learning Workflow

1. Frame the ML problem by looking at the science need
   a. Identify subproblems
2. Gather the data and do Data Munging/Wrangling
   a. Explore the data
   b. Clean data and prepare for the downstream ML models
3. Explore different models, perform V&V and shortlist promising candidates
4. Fine-tune shortlisted models, draw insights, and combine them together to form the final solution
5. Present your solution
   a. Say a story with the data
6. Deploy and monitor: Write a Paper; Publish Code etc

Deepak Subramani, deepakns@iisc.ac.in

Data Science Problem: Predict Habitat Suitability

Machine Learning for Asian Elephant Habitat Suitability Estimation in India



Predictors and Targets

Category	Variable	Source	Unit	Spatial resolution
Climatic	Monthly precipitation	https://www.worldclim.org	mm	2.5 minute
	Monthly minimum temperature	https://www.worldclim.org	°C	2.5 minute
	Monthly maximum temperature	https://www.worldclim.org	°C	2.5 minute
Topographic	Elevation above sea level	https://www.worldclim.org	m	30 arc-seconds
	Distance to rivers and water-bodies	Derived using QGIS with data downloaded from https://www.openstreetmap.org	m	
	Distance to roads	Derived using QGIS with data downloaded from https://www.openstreetmap.org	m	
	Land Use Land Cover (LULC)	https://bhuvan.nrsc.gov.in	categorical	30m
Vegetation related	Net Primary Productivity (NPP)	https://neo.sci.gsfc.nasa.gov	gC/m²/day	0.1 degrees
	Leaf Area Index (LAI)	https://neo.sci.gsfc.nasa.gov	m²/m²	0.1 degrees
	Normalized Difference Vegetation Index (NDVI)	https://neo.sci.gsfc.nasa.gov	Dimensionless	0.1 degrees

Input to the ML

X

Target of the ML

y

GBIF
Global Biodiversity Information Facility

Asian Elephant Presence Data

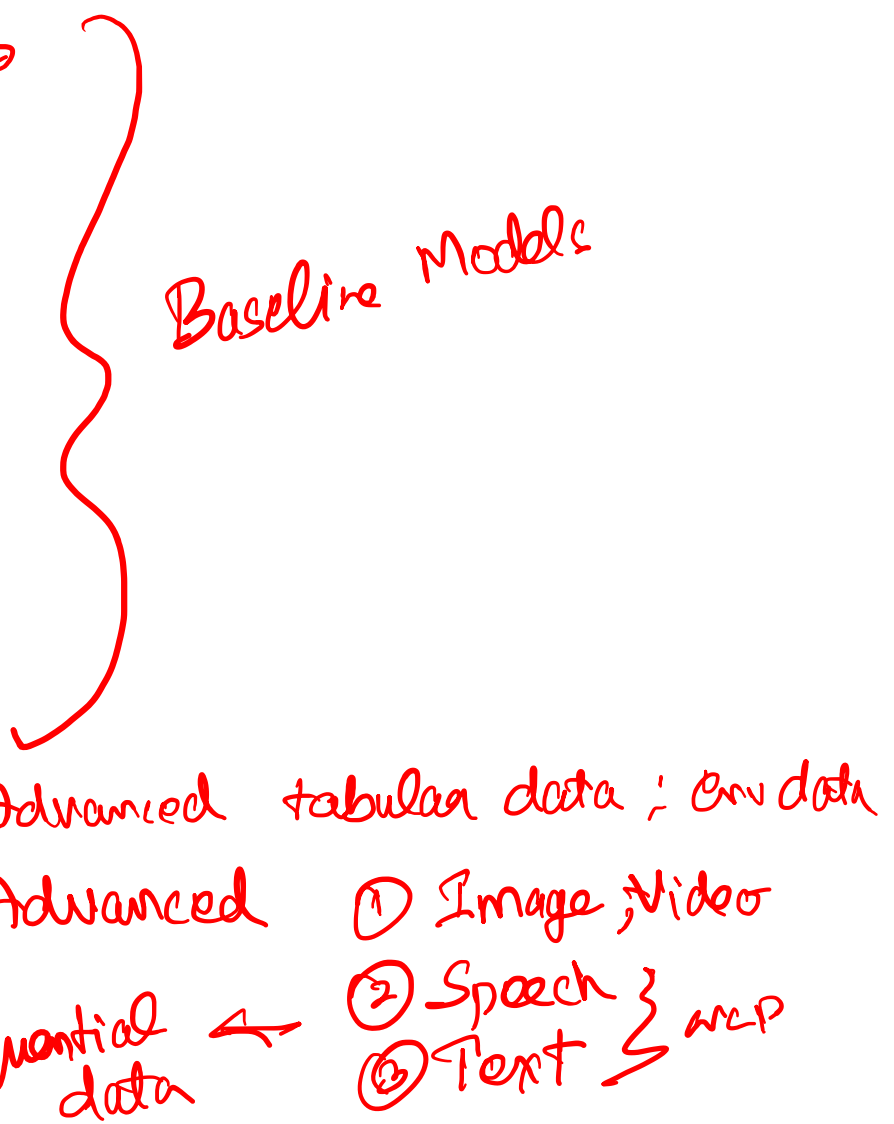
Paper: arXiv:2107.10478
2nd Place Award in InGARSS 2021

ML Problem – The view

θ

Step 3: Explore different ML Models

1. Linear Regression with Regularization
   - For regression task – predicting a continuous variable
2. Logistic Regression
   - For classification tasks alone
3. Support Vector Machines – Linear and Kernel
4. K-Nearest Neighbours
   - Instance-Based Method
5. Naïve Bayes
   - Simplest Bayesian Network Model
6. Decision Tree, Random Forests, XGBoost
7. Neural Networks



Deepak Subramani, deepakns@iisc.ac.in

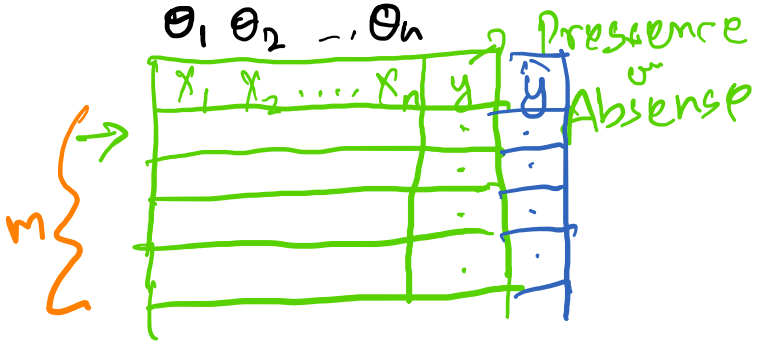
Classification Models

- Predict the value of a categorical variable in a supervised setting
- Use other variables as predictors

$$\hat{y} = h_\theta(X; \theta)$$

- Variable with a hat, $\hat{y}$ is usually the prediction from a model
- h is called the hypothesis, or the ML Model parametrized by θ
- θ can be calculated, estimated or ***learned*** from data by solving the optimization problem

1. Decision Variable
2. Objective Function
3. Constraint on the D.V.



$$\min_{\theta} \sum_{j=1}^{m} L(\hat{y}^{(j)}, y^{(j)})$$

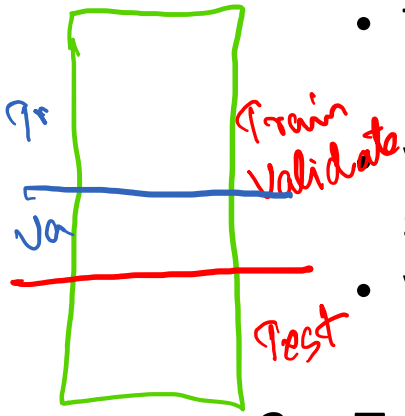
$$\min_{\theta} \sum_{j=1}^{m} L(\hat{y}^{(j)}, y^{(j)})$$

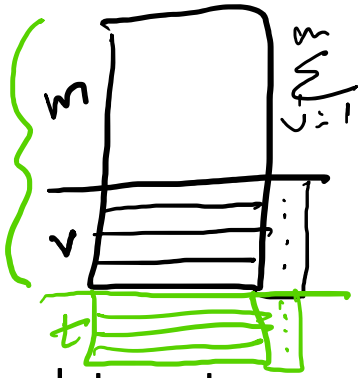
$$\min_{\theta} \sum_{i=1}^{m} L(h_\theta(x^{(i)}; \theta), y^{(i)})$$

Deepak Subramani, deepakns@iisc.ac.in

Training, Validation, Testing

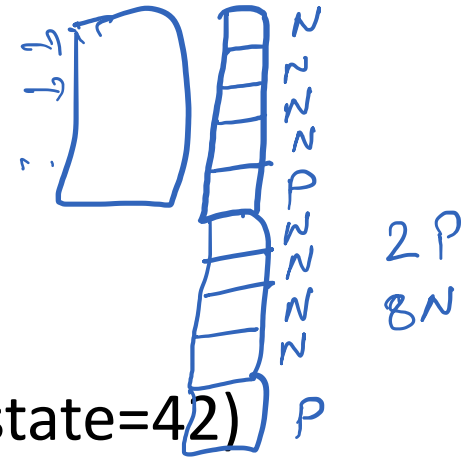
- In all settings, developing a ML model involves two stages
  1. Training-Validation
     - Training: Find the parameters of the ML model
       - Parameters are trained from data directly by an optimization algorithm
     - Validation is to see if parameters found during training are generalizing well to data not seen
     - Validation is used to tune the hyperparameters of the ML model
       - These are set by users and not directly computed by the optimization algorithm from data
  2. Testing
     - Testing is similar to validation, but the performance at test stage is not generally used to improve the model
     - Test data is completely unseen during model development





Train-Test Split – 2 Strategies

- We must separate the dataset to training and testing sets
  - Training set is used for building the model and testing set it used to evaluate model performance
  - The train set can be further split to training set and validation set
- Test set should not be used for model building purpose



Agnostic Random Split

```
from sklearn.model_selection import train_test_split
train_set, test_set = train_test_split(dataset, test_size=0.2,random_state=42)
```

Split to avoid sampling Bias

```
from sklearn.model_selection import train_test_split
split = train_test_split(dataset, test_size=0.2, random_state=42,
stratify=dataset[“label”])
```

Classification: Performance Measures

- Can we use plain accuracy?
- Let us design a classifier (i.e., a medical diagnostic test) for Covid-19.
  - The test is such that it simply return –ve for everyone
  - What will be the accuracy? – Greater than 90% as only 1% of the population currently has ever had the disease, and only about 5% of those testing are +ve
  - So is our classifier good? Surely >90% accuracy is good? No? Why?
- Accuracy is generally not enough for use as a performance measure for classifiers, especially for skewed datasets where some classes are more frequent than others (as in Covid-19)

Success Metrics, Validation of ML Models

- Classification
  - Accuracy
  - Recall
  - Precision
  - F1 Score (Jaccard Index)
  - Receiver Operator Characteristics
  - Cross Entropy

Success Metrics, Validation of ML Models

- Success Metrics are defined based on the task we perform
- Regression
  - Mean Squared Error and Root Mean Squared Error
  - Mean Absolute Error
  - Mean Relative Absolute Error
  - R^2 for linear regression problems [Very common in statistics, and early days]
  - Advanced use: Maximum error, MSE only for some part of the data, etc

Deepak Subramani, deepakns@iisc.ac.in

Classification Models

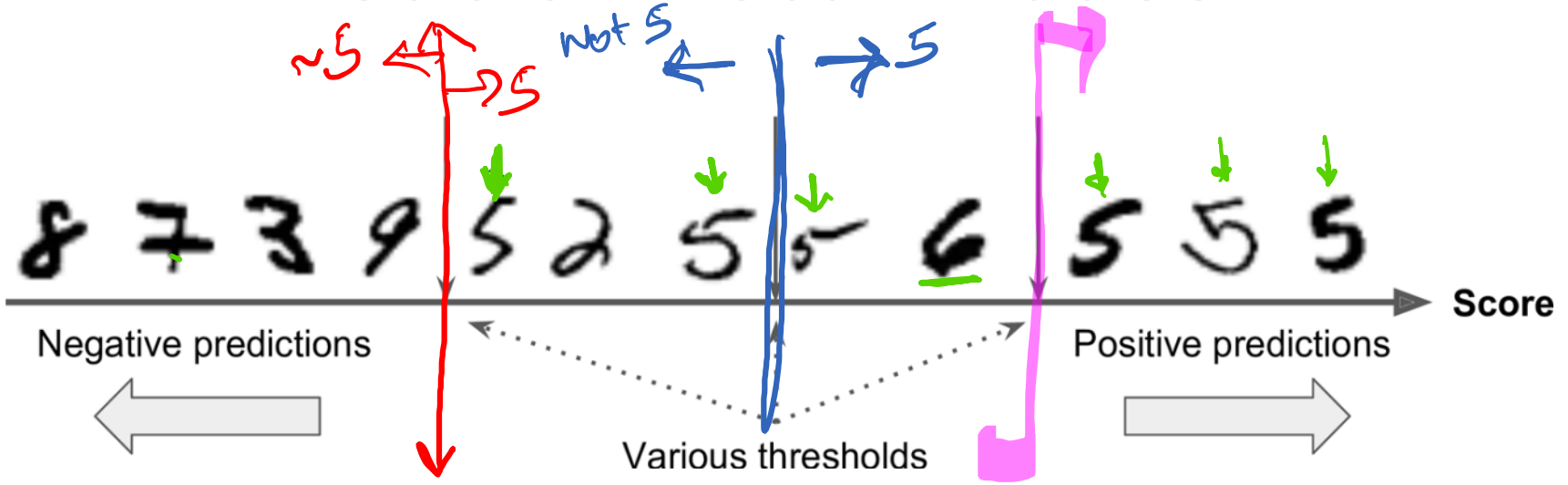
- Other than Regression, where we predict continuous values, Classification, where we predict classes is an important supervised learning task
- Classification models are data-driven models that enable us to distinguish which class a data point belongs to.
- Binary Classifier: A data-driven model that identifies if a particular combination of attributes (i.e., data) belongs to one class (1) or not (0)
- Evaluating a classifier is the tricky part which we will dive deep into first.

Confusion Matrix

	True Label A	True Label ~A
Pred. Label A	True Positive	False Positive
Pred. Label ~A	False Negative	True Negative

- Precision = $\frac{TP}{TP+FP}$;
  - Among all prediction of Label A, how many are actually Label A
  - Trivial 100% Precision – Make only one Prediction of Label A, and ensure that it is correct. Then TP=1, FP=0, and Precision=1
- Recall = $\frac{TP}{TP+FN}$;
  - Of all true Label A, how many does our classifier predict as Label A
  - Combined with Precision, we now have a good sense of the goodness of our classifier
- Typically, we want high precision and high recall
- F1 Score = $\frac{2}{\frac{1}{precision}+\frac{1}{recall}} = \frac{TP}{TP+\frac{FN+FP}{2}}$
- F1 Score is the harmonic mean of precision and recall. Hence F1 will be high only if precision and recall are high

Precision Recall Tradeoff



	True 5	True ~5
Pred 5	4 TP	1 FP
Pred ~5	2 FN	5 TN

$$\text{Precision} = \frac{4}{4+1} = \frac{4}{5} = 0.8$$

$$\text{Recall} = \frac{4}{4+2} = \frac{4}{6} = 0.67$$

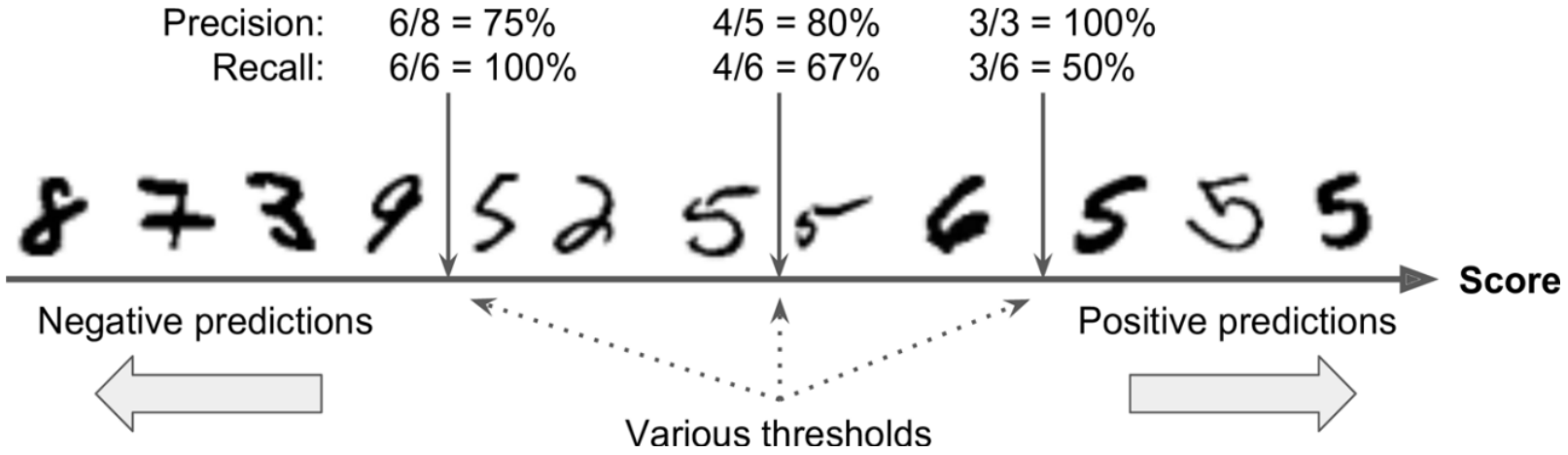
	T 5	T ~5
P 5	6	2
P ~5	0	4

$$\text{Pre} = \frac{6}{6+2} = \frac{3}{4}$$

$$\text{Rec} = 1$$

Deepak Subramani, deepakns@iisc.ac.in

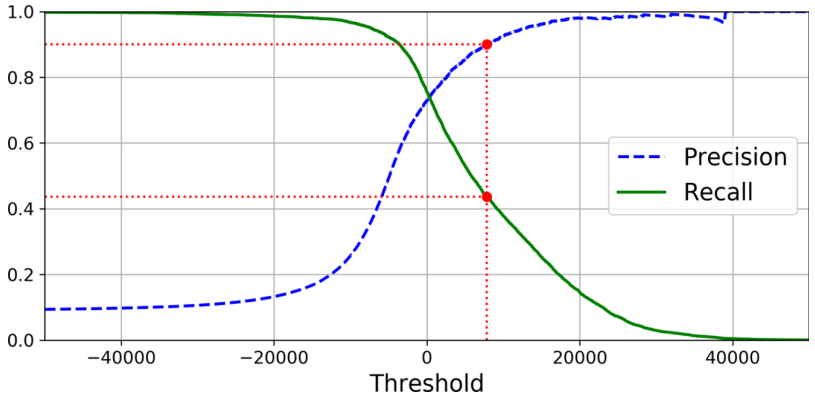
Precision-Recall Tradeoff



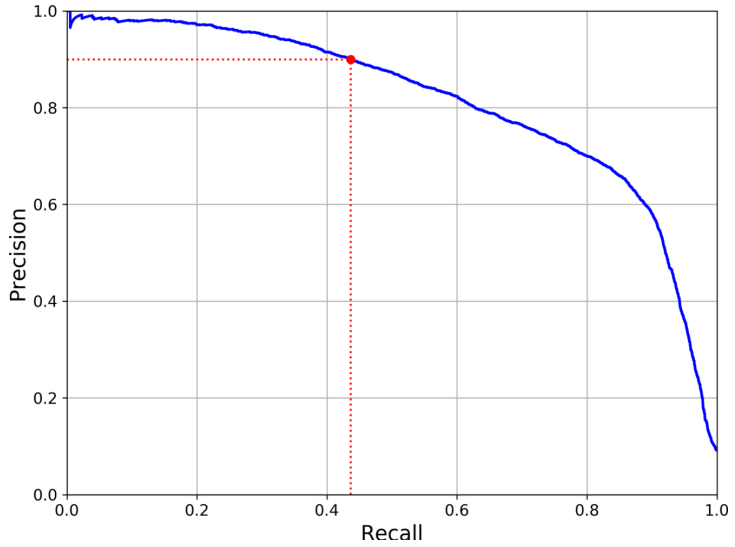
- Classification is based on a score calculation and a threshold set
- Based on where the threshold is placed, precision increases at the expense of recall, and vice versa
- While F1 Score is good, in some applications, we prefer high recall even at the expense of low precision
  - For example, fraud detection. We want to catch all the fraud cases, even if some non fraud cases are also flagged.

Precision-Recall Tradeoff

	True Label A	True Label ~A
Pred. Label A	True Positive	False Positive
Pred. Label ~A	False Negative	True Negative



- If you label almost all as A (low threshold), then recall is high, but precision is poor
- If you label almost all as ~A (high threshold), then precision is potentially high (or Not defined), but recall is nearly zero
- If TP=0, both are zero
- For different thresholds, we usually plot both precision and recall vs threshold.
- We can also plot Precision vs Recall
- Choose the right threshold from the above curve and settle for a Precision-Recall tradeoff



Deepak Subramani, deepakns@iisc.ac.in

Plots from Geron Chapter 3

Other Definitions

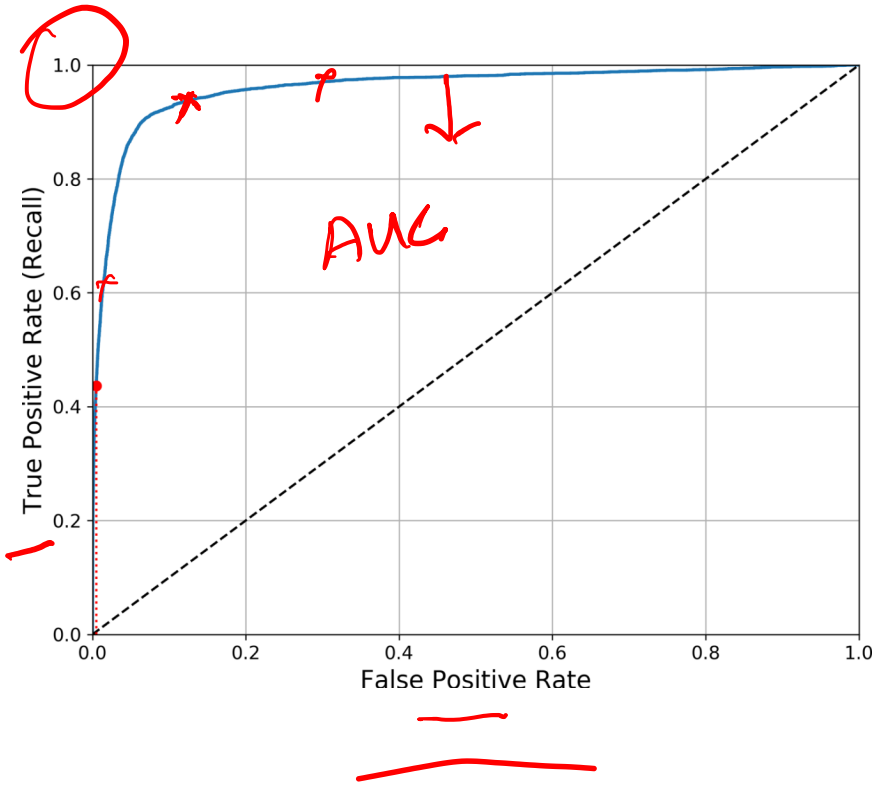
		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

Confusion Matrix; Wikipedia

Deepak Subramani, deepakns@iisc.ac.in

Receiver Operating Characteristic (ROC) Curve

- True Positive Rate (TPR) – Recall or Sensitivity
- True Negative Rate (TNR) – Specificity – Ratio of negative instances correctly classified as negative
- False Positive Rate = 1 – True Negative Rate
- ROC plot is sensitivity vs 1-specificity, i.e., TPR vs FPR
- Compute TPR and FPR for different thresholds and plot it
- The ROC must be as much away from the 45-degree line as possible
- Calculate the Area Under the Curve (AUC) for quantifying goodness of the classifier
- AUC=1 is a perfect classifier, and AUC=0.5 for a random classifier



Deepak Subramani, deepakns@iisc.ac.in

PR vs ROC – Which to Use?

- As a rule of thumb, prefer the PR curve whenever
  - the positive class is rare or
  - when you care more about the false positives than the false negatives.
- Otherwise, use the ROC curve.

Main Challenges of ML

- Insufficient Training Data
- Non-representative/Biased/Skewed Training Data
- Poor Quality/Error Prone Data (Missing entries etc)

Data

- Overfit or Underfit

Deepak Subramani, deepakns@iisc.ac.in

Audience Poll

- What is a error metric for regression
  - F1 Score, RMSE, Precision, Recall
- F1 Score is the Arithmetic Mean of Precision and Recall
  - True, False
- When labels have a skewed distribution which sampling strategy must be used for train-test split?
  - Uniform Random Sampling, Stratified Sampling, Choose first r% of data
- Which of the following is true?
  - When precision increases, recall must decrease
  - When precision and recall increase, F1 score decreases
  - A classifier with high area under ROC will have high recall
  - Recall is also called false positive rate