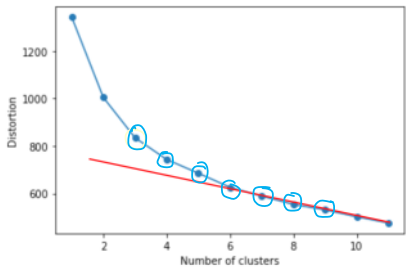
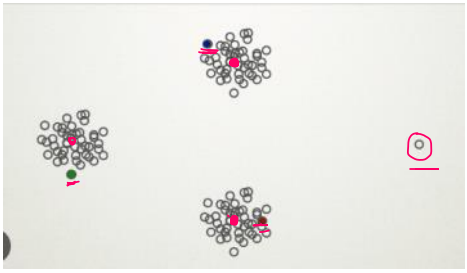
## Lecture: 131

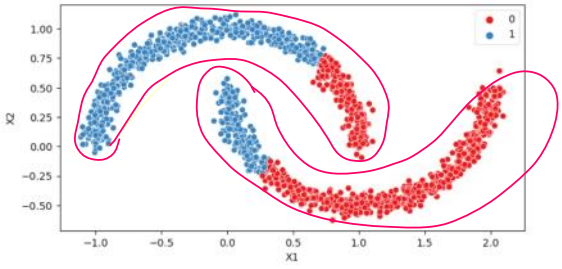
DBSCAN এটা Clustering Algorithm। এর আগে আমরা K-mean Clustering করেছি। -কেন DBSCAN পড়বো? এর আগে K-mean এর কিছু limitation সম্পর্কে জেনেছি।

Why DBSCAN?

16 December 2023 15:10






## Disadvantage of K-Mean Clustering:

(i) -কয়টা cluster হবে সেইটা বলে দিতে হয়। Elbow-method দিয়ে যখন -কয়টা cluster হবে সেইটা বের করতে যাবো গ্রাফ টপকের মতো আসলে -কয়টা cluster হবে সেইটা অনুমান করা কষ্টসাধ্য।

(ii) Outliers থাকলে K-mean clustering ভালো perform করে না।

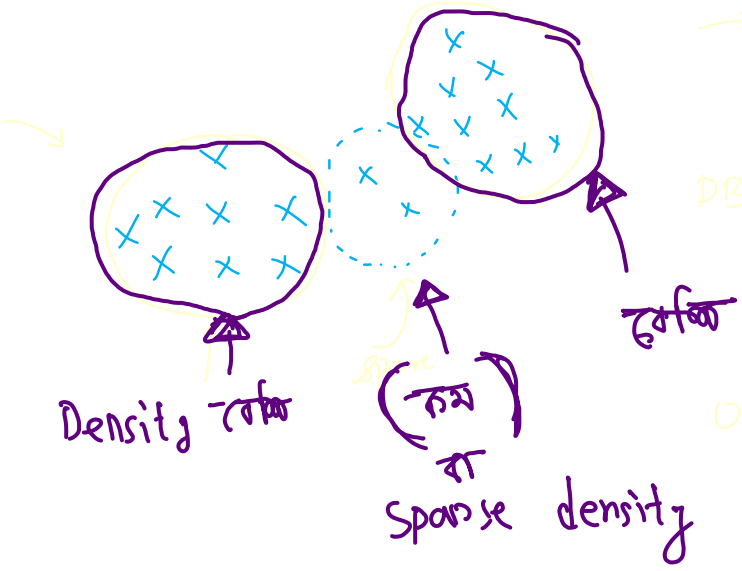
(iii) K-Mean দিয়ে বানানো clusterগুলো অনেকটা গোল shape ই হয়। যদি, cluster গুলো complex হয়, যেমন ডিমের কুসুমের মতো, তাহলে K-mean ভালো perform করতে পারে না।

গোল shape হওয়ার কারণ, K-mean হচ্ছে centroid based clustering algorithm.

কিন্তু DBSCAN হচ্ছে Density Based Clustering Algorithm

# What is Density Based Clustering

16 December 2023 15:12




Sparse density বা কম density কে টিপার ভিত্তি করে, DBSCAN আন্দাজ করতে পারে, এখানে, ২ টির মতো cluster হতে পারে।

তাই,

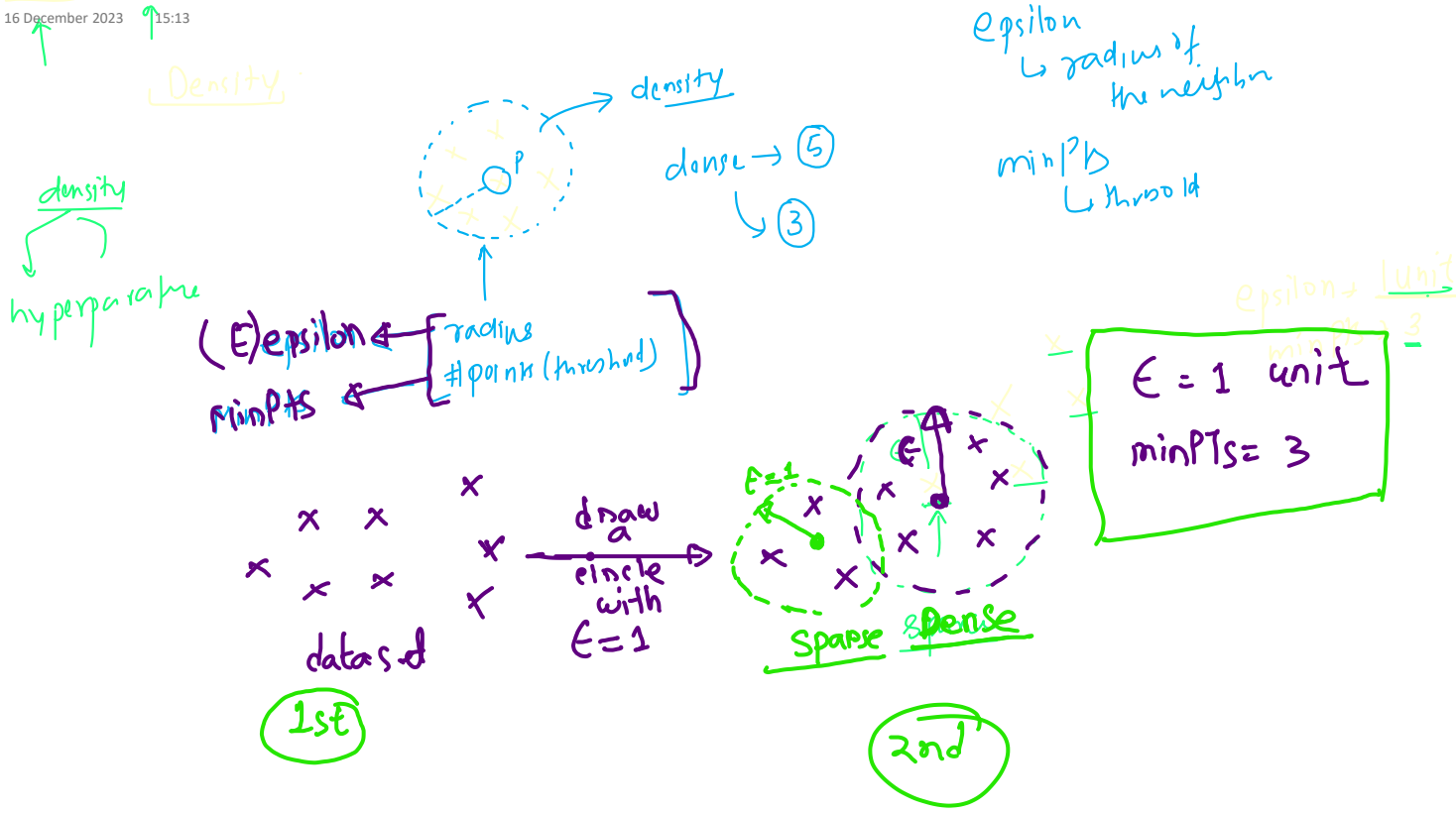
DBSCAN → Density Based Spatial Clustering of Application with Noise

Density Based আরেকটা Algorithm আছে যার নাম OPTICS

# Hyperparameter (MinPts & Epsilon) in DBSCAN

MinPts & Epsilon

16 December 2023 15:13




1st ও আমাদের dataset দেওয়া আছে। আর, $\epsilon = 1$ and minPTS = 3 দেওয়া আছে। প্রথমে dataset (Epsilon, $\epsilon = 1$) এর সমান radius নিয়ে circle করা হবে, তারপর যদি circle এর মধ্যে minPTS এর সমান বা বেশি point থাকে তাহলে তাকে Dense বলবো আর না থাকলে Sparse বলবো।

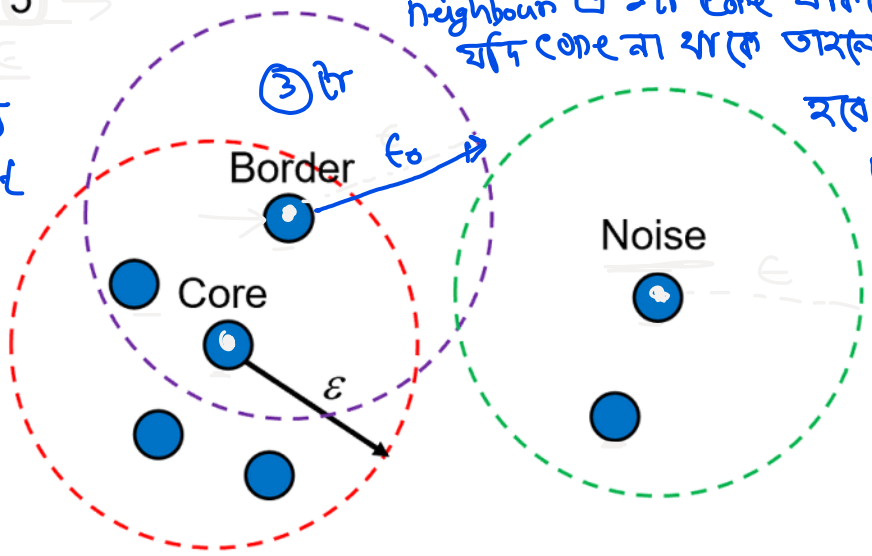
# 3 types of point in DBSCAN

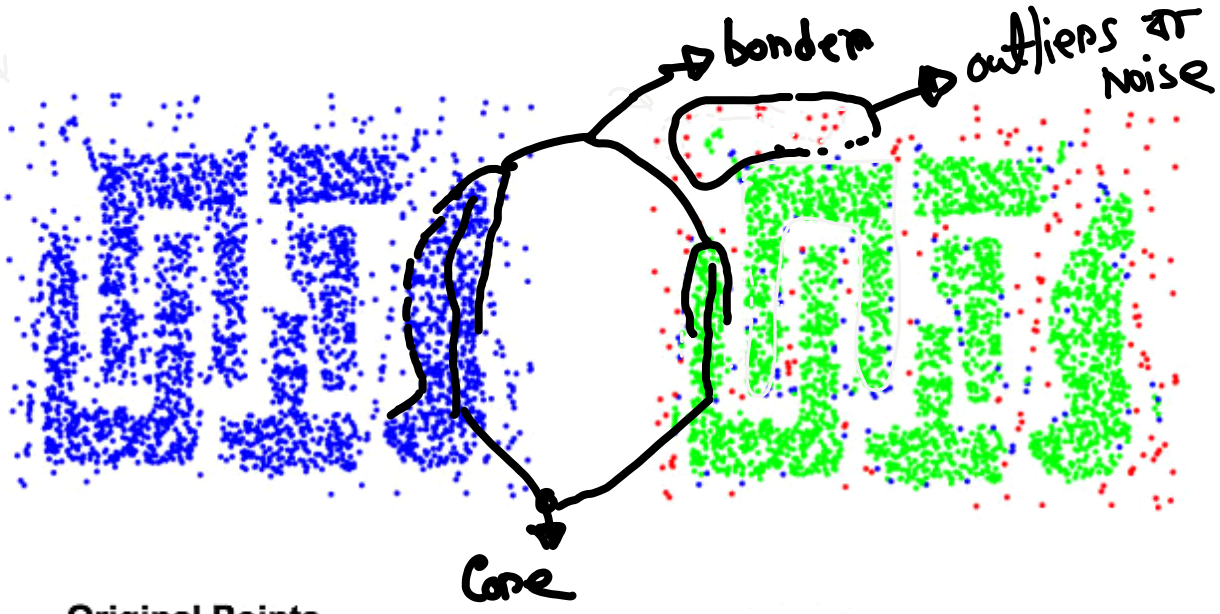
## Core Points, Border Points & Noise Points

16 December 2023 15:13

**Core:**
কোন data point এর জন্য
ε এর neighbour data point
এর সংখ্যা minPts এর সমান
বা বেশি হলে তাকে
core point বলে।
↳ মানে হচ্ছে circle বানাবো

**Border:** ε এর neighbour এর সমস, total data point MinPts এর থেকে কম। যেমনে, 3টা < 5টা। আর, ε এর neighbour এ 2টা Core থাকবে। আর যদি Core না থাকে তাহলে সেটা হবে Noise point.




Original Points

Point types: core, border and noise

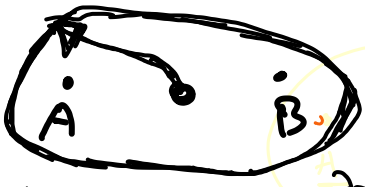
Eps = 10, MinPts = 4

# Density Connected Points

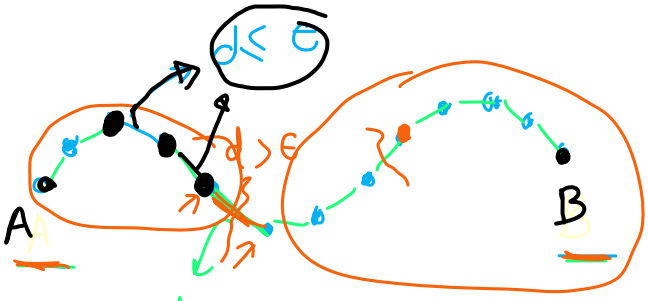
16 December 2023 15:13

Density Connected Points

এখন, A B ২টা point Density connected সেইটা কিভাবে বুঝবো?




simply mean করে, ২টা point A,B Density Connected হলে তাদেরকে ব্যবহার করে আমরা ১টা cluster এর মধ্যে ফেলতে পারি।




যদি, A থেকে B এর মধ্যে এমন point থাকে যাদের দূরত্ব ∈ এর থেকে কম।

d < ϵ

(মধ্যবর্তী)

[ Now let's learn the DBSCAN Algorithm ]

# DBSCAN Algorithm

16 December 2023 15:14

**step0** - We decide the value of E & minPTS.

**Step 1** - Identify all points as either core point, border point or noise point

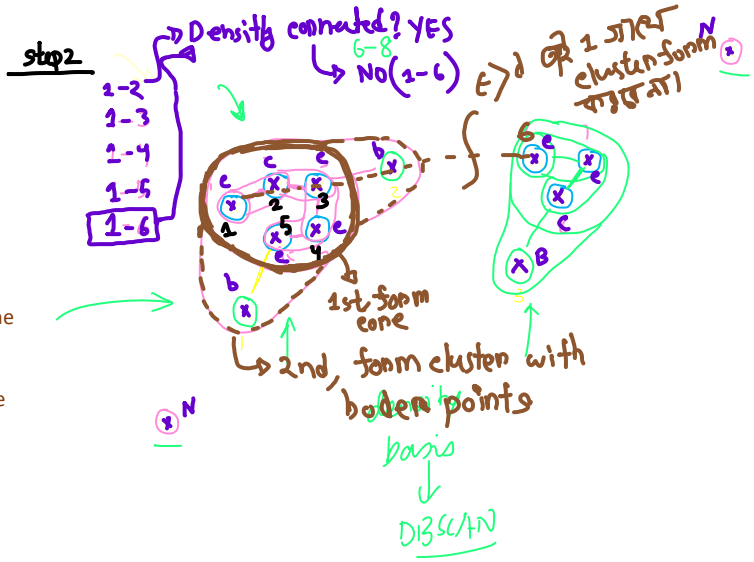
**Step 2** - For all of the unclustered core points

**Step 2a** - Create a new cluster

**Step 2b** - add all the points that are unclustered and density connected to the current point into this cluster

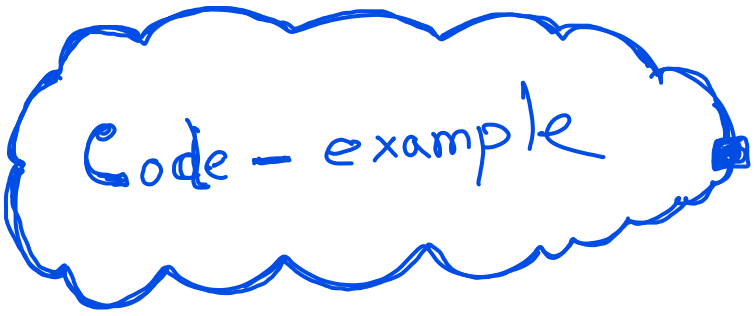
**Step 3** - For each unclustered border point assign it to the cluster of nearest core point

**Step 4** - Leave all the noise points as it is.



# Code

16 December 2023 15:14



Limitations

16 December 2023 15:14

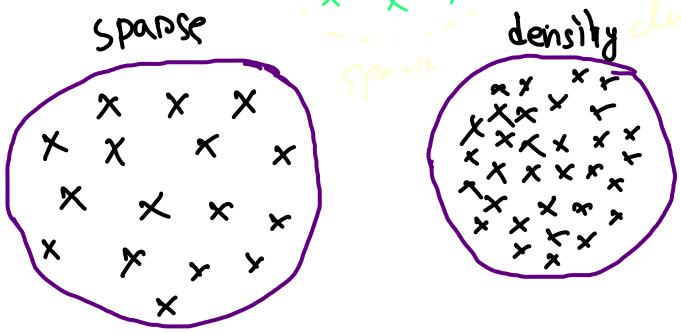
# (Advantage & Disadvantage of DBSCAN)

**Advantage** → তাই anolomy detection এ ব্যবহার করতে পারি।

1. Robust to outliers
2. No need to specify clusters
3. Can find arbitrary shaped clusters → যেমন ধরনেরই cluster দেয় না কেন কাজ করবে।
4. Only 2 hyperparameters to tune

**Disadvantage**

1. Sensitivity to hyperparameters
2. Difficulty with varying density clusters
3. Does not predict




এখানে, ২টা cluster আছে, কিন্তু ১টার density বেশি, অন্যটা কম, কিন্তু, আমাদের $\epsilon$ এর মান তো একটাই হবে তাই cluster এর খোঁজে। এই ধরনের data তে DBSCAN ভালো perform করে না।

***

সবচেয়ে মজার বিষয় হচ্ছে, DBSCAN এ predict বলে কোন fn নেই। কারণ, DBSCAN নতুন Data এর উপর prediction করতে পারে না।

# Visualization

16 December 2023 15:14


Imbalanced Data in ML

Lecture: 132

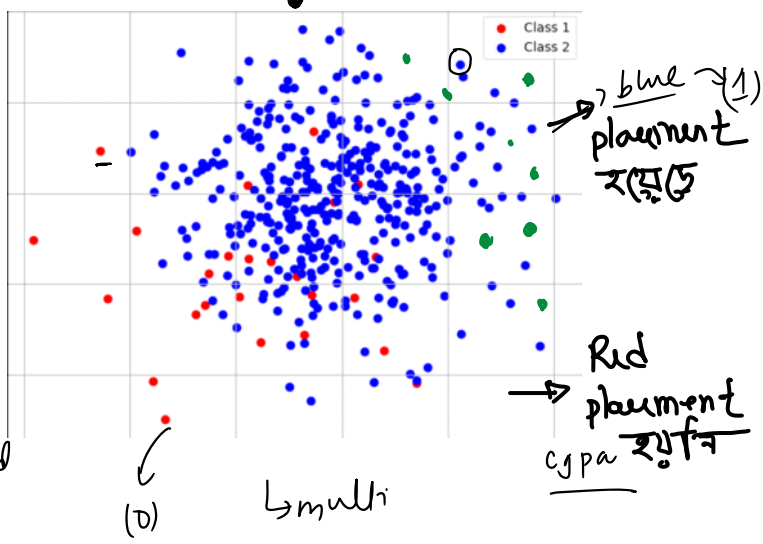
# What is Imbalanced Data?

02 May 2024 16:12

500 students → model → {cgpa, iq} → 0, 1 (binary)

| CGPA | IQ | Placement |
|---|---|---|
| 5.49 | 89 | 0 |
| 7.15 | 137 | 1 |
| 6.03 | 96 | 0 |
| 5.45 | 138 | 1 |
| 4.24 | 131 | 1 |
| 6.46 | 87 | 0 |
| 4.38 | 75 | 1 |
| 8.92 | 127 | 1 |
| 9.64 | 122 | 1 |
| 3.83 | 59 | 1 |

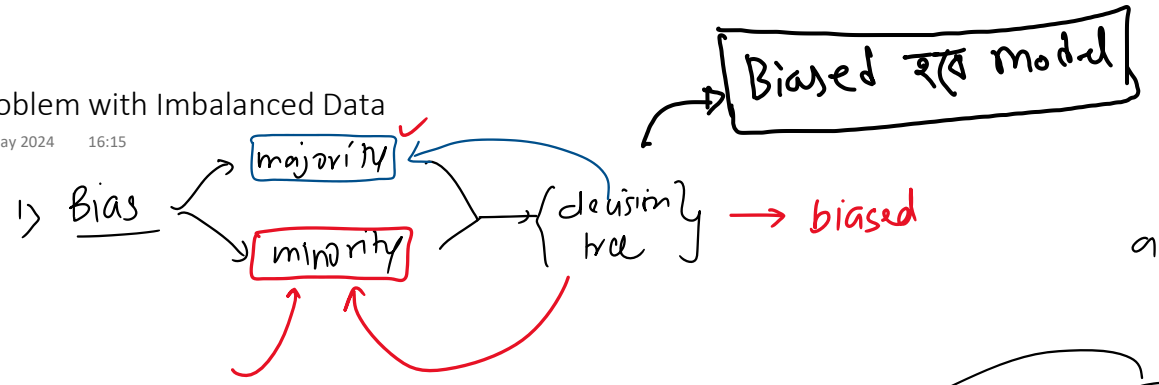
450(1) 50(0) imbalanced




imbalance data বলতে এমন data কে বুঝায়, for example উপরের ক্ষেত্রে placement হয়েছে ধরে (1) আর না হয় ধরে না (0)। এখানে, 1 এর থেকে 0 এর সংখ্যা অনেক কম, তাই এই dataset দিয়ে model train করলে, (1) দিকে biased model থাকবে। এই, ধরনের dataset -কে imbalance dataset বলে।

# Problem with Imbalanced Data

02 May 2024 16:15



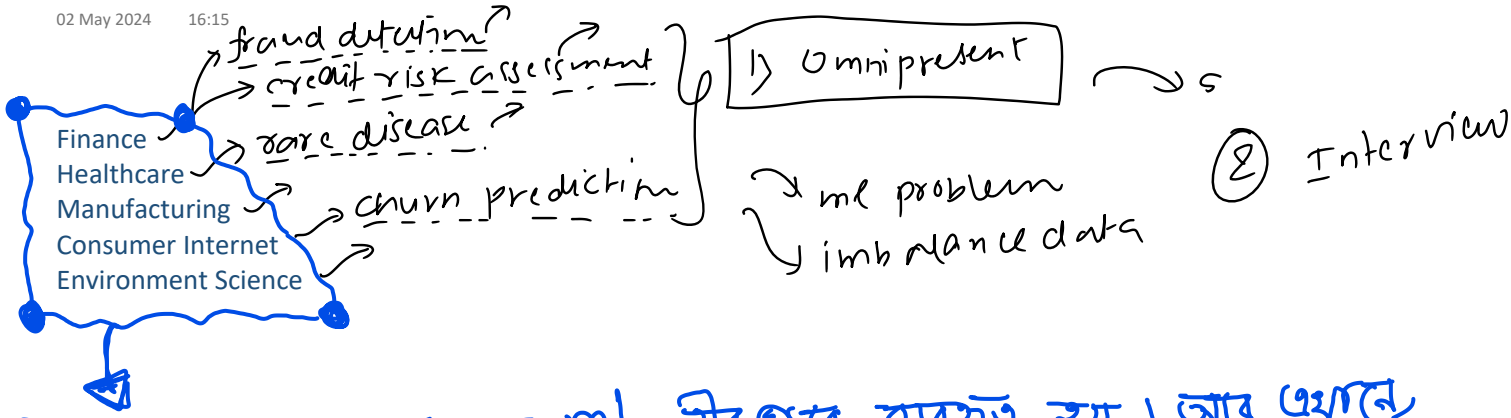

যেখানে, Yes বেশি
No কম, test করলে,
1000 থেকে 900 বার might ans দিবে
আর 100 বার wrong, তাহলে, accuracy
90% । কিন্তু, F1, Recall, Precision
দেখলে বুঝবো এর actual অবস্থা কী?

# Code Example #

# (কেন imbalanced Data গুরুত্বপূর্ণ ML এ)? #

# Why studying Imbalanced Data is important?

02 May 2024 16:15




এই field গুলোতে এখনো ml প্রবলেমে ব্যবহৃত হয়। আর এখানে imbalanced data থাকবেই।

- Frad Detetion
- Credic Risk Assesment
- Rare Disease
- CHun Prediction
- Earth Equick

এখানে, হাজার জনের মধ্যে ফ্রড থাকবে কম। তাই লোন load দিলে কেউ ফ্রড করতে পারবে না এমন লোকজন কম। Rare disease অনেক কম মানুষের হবে। খুব কম সংখ্যক লোকই এটার platform হয়ে। আর, কোন এক region এ earth quick কম হয়। অর্থাৎ, এখানে সকল data হচ্ছে imbalanced.

→ Interview তে ask করে অনেক।

## Some Technique to deal with imbalanced data.

(i) Undersampling

(ii) Oversampling

(iii) SMOTE

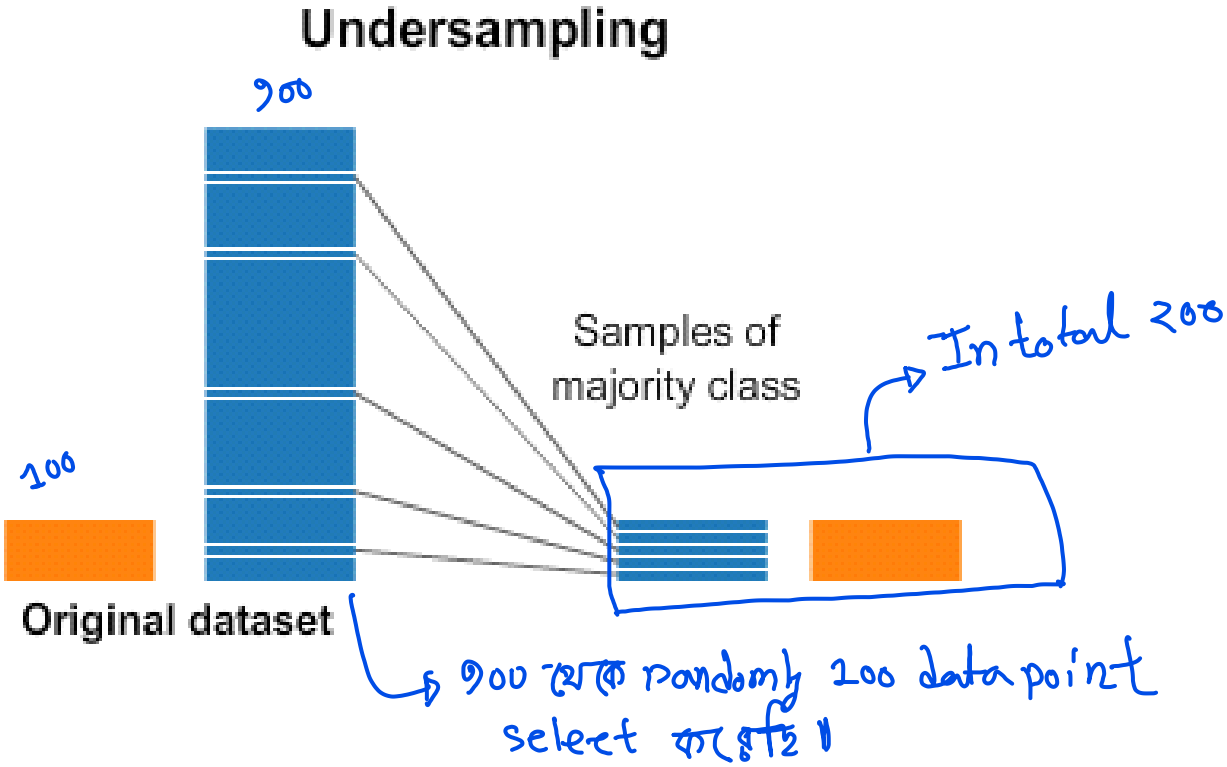
(iv) Ensamble Learning

(v) Cost Sensitive Learning

(vi) Others

# 1. Undersampling

02 May 2024 16:17




Code example

**Advantages :**

1. Reduction in bias
2. Faster training

**Disadvantages :**

1. Information loss leading to underfitting
2. Sampling Bias → Randomly data sample pick করার কারণে Bias হওয়ার সম্ভাবনা।

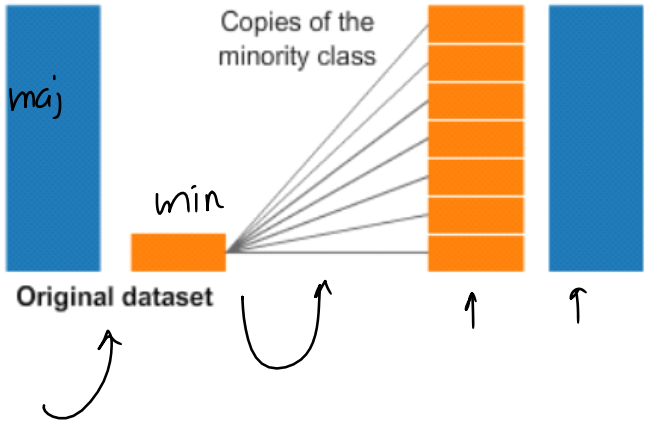
## 2. Oversampling

03 May 2024 01:13

→ [যেখানে, datapoint add করি।]

Random

**Oversampling**




maj 900

min 100

balance

900 duplicate

800 → duplicating

বাকী 800data duplicate করার মাধ্যমে achieve করি

# Code Example #

Advantage

1. Reduced Bias ✓

Disadvantage

1. Increased size
2. Duplication of data may cause overfitting

ml also

simp → overfitting

↓

decision tree ও ROS → overfitting

→ like, DT এর ক্ষেত্রে Oversampling করলে, overfitting হওয়ার চেষ্টা।

## Synthetic Minority Over-sampling Technique:

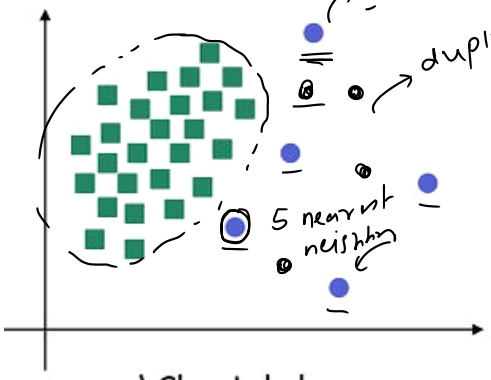
3. SMOTE

popular → Oversampling → min → upsample

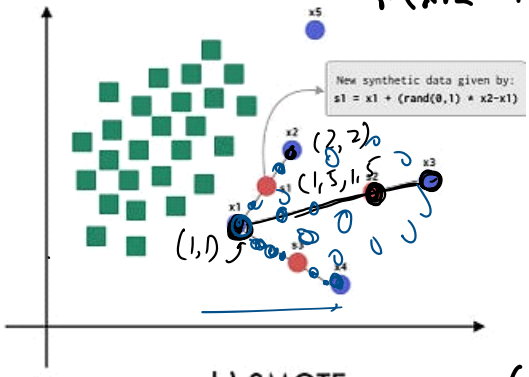
oversampling → duplicate

interpolation

এখানে, minvalue গুলোকে upsampling করি, interpolation এর মাধ্যমে over sampling এ যেখানে, duplicate করানো হয়।




a) Class Imbalance




b) SMOTE

Algorithm:

5 KNN

- Train a KNN on minority class observations - find each observation's 5 closest neighbours

**To create the new synthetic data:**

- Select examples from the minority class at random
- Select a neighbour of each example at random (for the interpolation)
- Extract a random number between 0 and 1
- Calculate the new examples as = original sample - factor * (original sample - neighbour)
- The final dataset consists of the original dataset + the newly created examples

(0-1) মধ্যে, 0.5 ধরেছি

sample − factor × [sample − neighbor]

[1,1] − 0.5 [(1,1) − (2,2)]

(1,1) − 0.5 [(−1,−1)]

(1,1) − (−0.5, −0.5)

= 2 [1.5, 1.5]

Data generate করতে থাকি যতক্ষণ না পর্যন্ত both Sample সমান হচ্ছে।

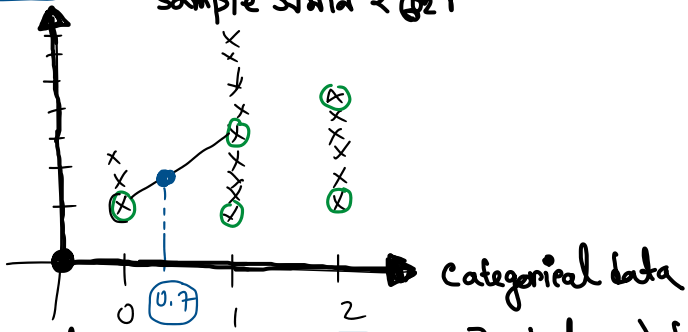
#Code example #

0 → K → K=1

K=5

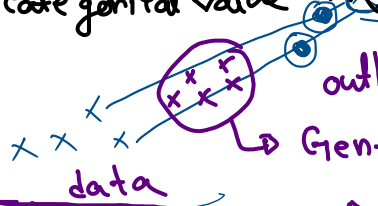
Disadvantages

1. Does Not Handle Categorical Data Well
2. Computational Complexity (KNN-এর কারণে)
3. Dependency on the Choice of Neighbors
4. Sensitive to Outliers
5. Balance Achieved May Not Reflect True Nature




Interpolation এখানে (0~1) এর মাঝে কোন data point Generate হবে এর জন্য কোন categorical value থাকবে না।

KNN এ, K=1 দিলে data point কম generate হবে, K এর মান বেশি দিলে data point বেশি generate হবে। K এর optimum value বের করা একটু কঠিন, তবে সাধারণত মান ৫ ধরে নেয়া যায়।

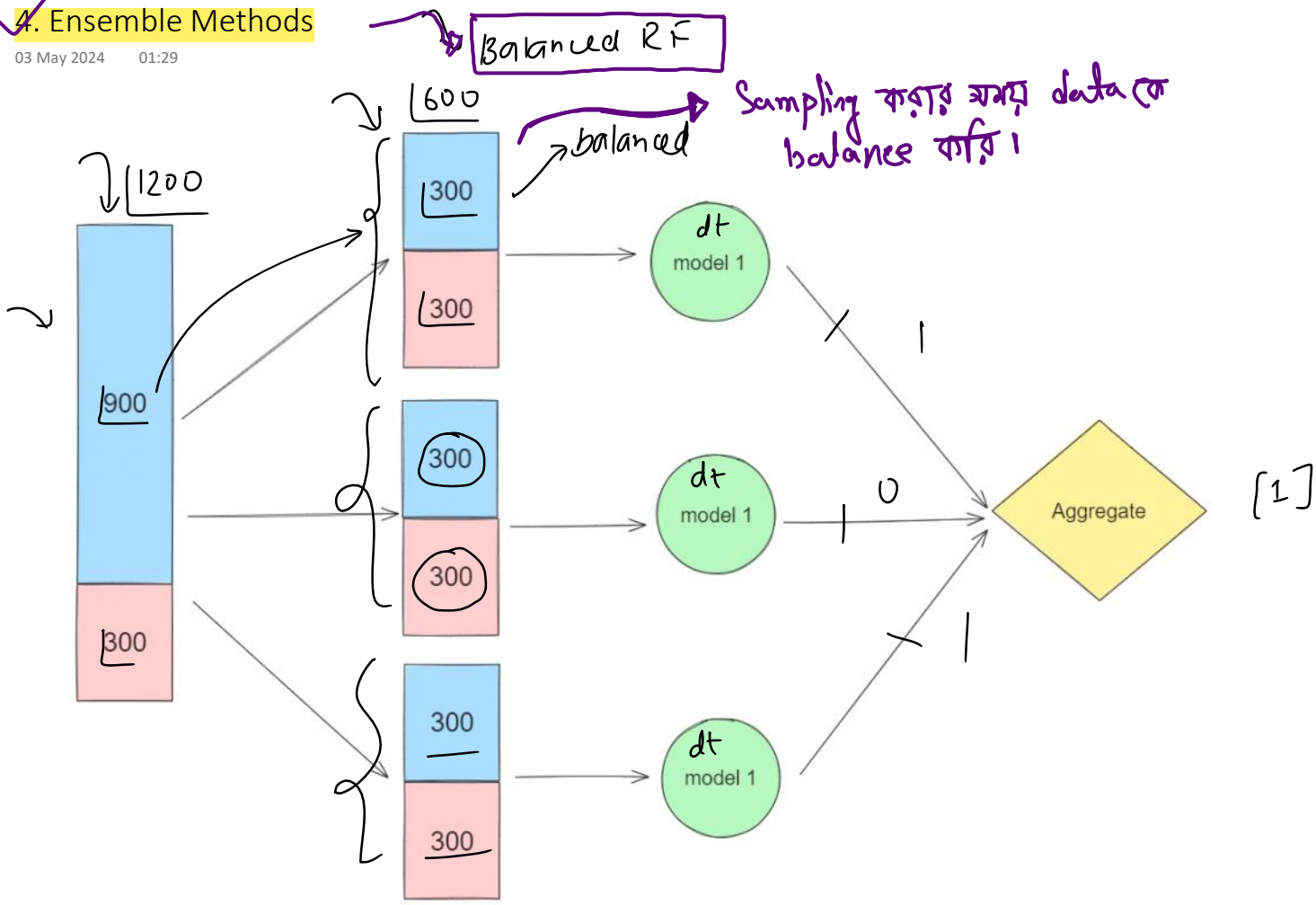


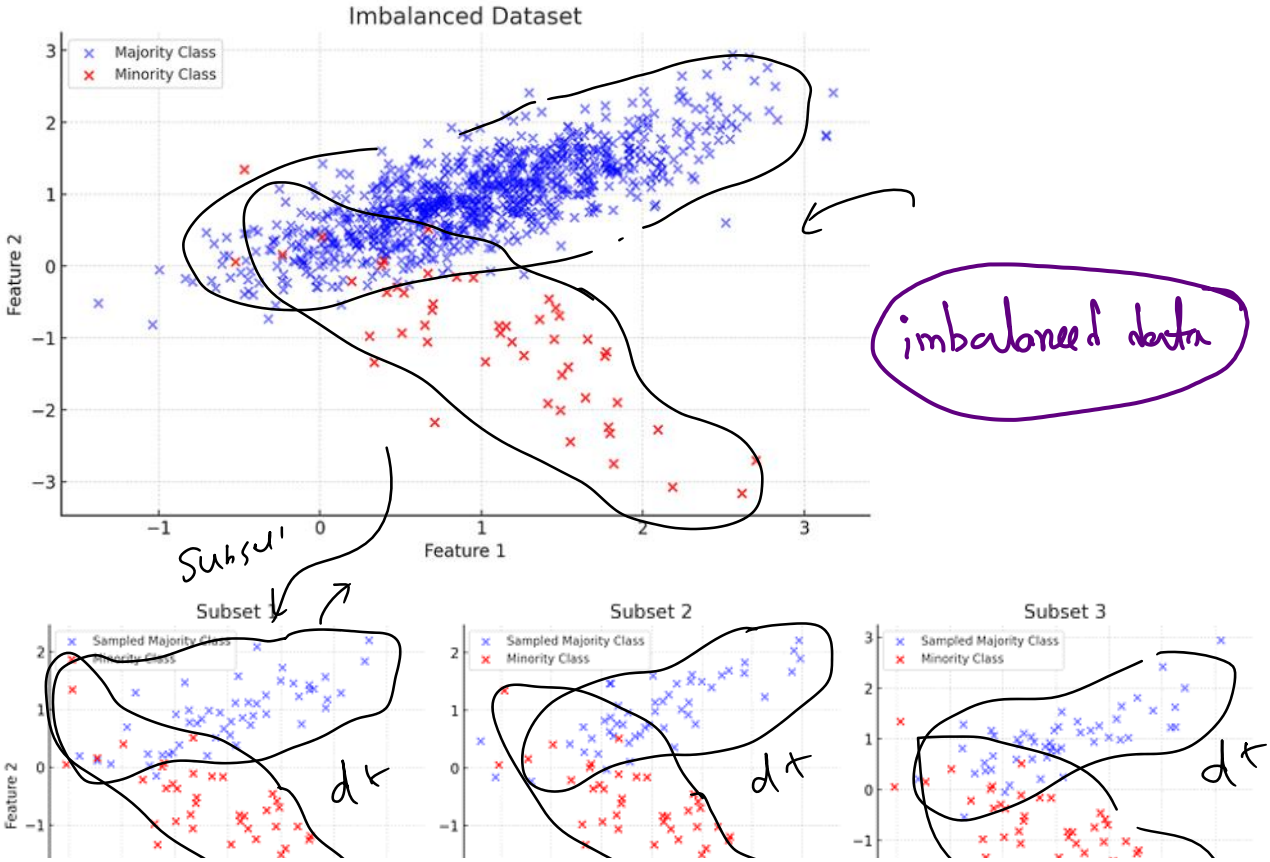

Outliers এর কাছে new generated Synthetic point গুলোর কারণে perform খারাপ হবে।

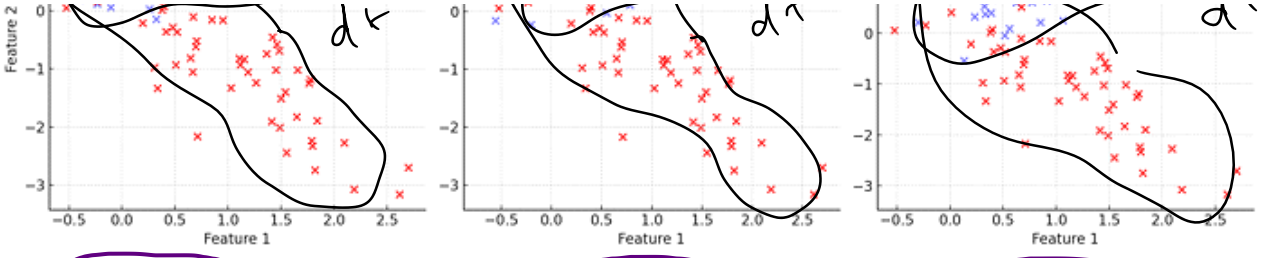
→ Balance achieve হলেও, এটা বাস্তবসম্মত নাও হতে পারে।

# 4. Ensemble Methods

03 May 2024 01:29






Sample-1 Sample-2 Sample-3

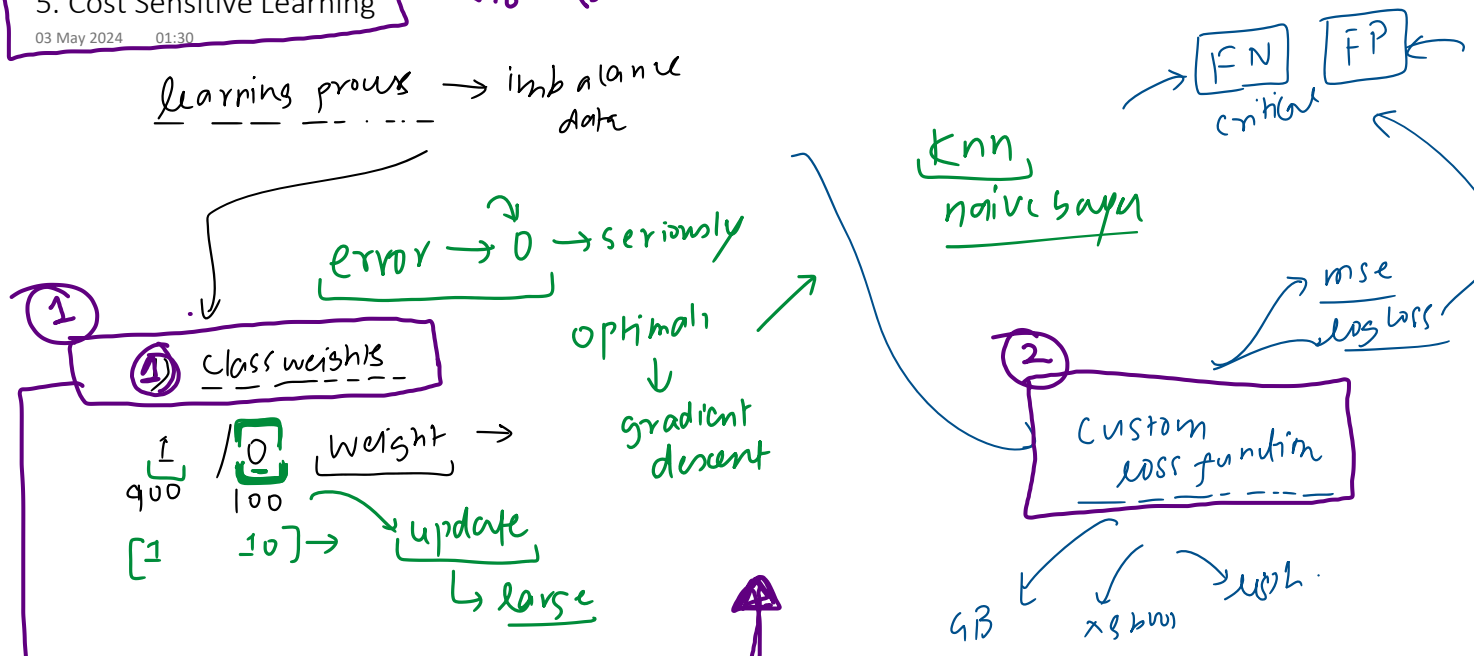
↳ অর্থাৎ, ensamble learning এর যে algorithm গুলোতে data কে sample করি, যেখানে আমরা imbalance data কে handle করতে পারবো।

# 5. Cost Sensitive Learning

03 May 2024 01:30

→ model এর tranning process এ একটু fog chance করি যাতে করে imbalance data handle হয়।

learning process → imbalance data




আমরা, class weight দিতে পারি।
Majority class কে কম আর minority class কে বেশি weightage দিতে পারি।
বা, tranning process এ যখন minority class এর উপর error আসবে, সেইটাকে একটু বেশি গুরুত্বসহকারে দেখবে।
sk-learn এ অনেক algorithm এ, class_weight নাম parameter আছে।

like: (i) LOR
(ii) SVM
(iii) DT

But, KNN, Naive bayes এই গুলোতে থাকে না।

② custom loss fn design করে। যেমন: GB, xgboost এ custom loss fn ব্যবহার করা যায়।

# List of all the techniques

03 May 2024 15:06

Imbalance learn library এর Documentation follow করবো। সেখানে আরো অনেক technique আছে imbalance data handle করার জন্য।