Agenda:
1. ROC
2. PR Curve — AUC
3. KNN
4. Geometric Intuition
5. Time Algorithm
6. SCIKIT Implementation
7. Time Space Complexity
8. Bias Variance Tradeoff

ROC
(Receiver Operating Characteristic) — ROC

→ electronics & Radio engineering

Context: Binary classification (2/0)

measure the goodness of model ← AUC-ROC, AUC

Step

x	y
x^1	1
x^2	0
x^3	0
x^4	1
x^5	1
x^6	0
x^7	0
x^8	1
x^9	1
x^{10}	0

→ Model ($x^{(i)}$)

$\hat{y}$: 0.9, 0.7, 0.6, 0.4, 0.3, 0.2, 0.1, 0.0

Scored data

(1) Sort $\hat{y}$ → hot, $\hat{y}$ decs. order. $O(n \log n)$

(2) For each $\hat{y}$ use as a threshold

(a) $T_1 = 0.95$ → if $\hat{y}^{(i)} \geq 0.95 \rightarrow 1$, else 0

Actual & Yhat $\rightarrow TPR_1, FPR_1$

(b) $T_2 = 0.85$ → Get Yhat if $\hat{y}^{(i)} \geq 0.85 \rightarrow 1$, else 0

Actual & Yhat $\rightarrow TPR_2, FPR_2$

$T_n = 0.01$ → TPR_n, FPR_n

x	y	Yhat$_1$	Yhat$_2$	...	Yhat$_n$
x^1	1	1	1		
x^2	0	0	0		
x^3	0	0	0		
x^4	1	0	1		
x^5	1	0	1		
x^6	0	0	0		
x^7	0	0	0		
x^8	1	0	1		
x^9	1	0	1		
x^{10}	0	0	0		

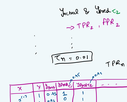
TPR_1, FPR_1 ... TPR_n, FPR_n

n-Pairs of (TPR_i, FPR_i)

TPR	FPR
0.1	0.0
0.2	0.1
0.3	0.1
0.4	0.2
0.5	0.3
0.6	0.4
0.7	0.5
0.8	0.6
0.9	0.7
1.0	0.8

Step

AUC-ROC of (Ideal Model) = 1



Model A Vs. Model B

Model B > Model A

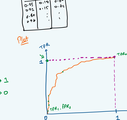
Understand 2 ff.
PR (Precision) F1 Score
(certain $t = 0.5$)

ROC → all possible threshold.

Random Model AUC ⇒ 0.5

Ex.

Group	Yhat
1	1
1	0
1	0
1	1
1	0
0	1
0	0
0	1
0	0
0	0

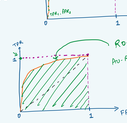


Properties of AUC-ROC
1. data is highly imbalanced, doesn't work.
2. AUC-ROC doesn't depend on $\hat{y}^{(i)}$ value. — rather depends on the order.

$\hat{y}^{(i)}$	$y_{(i)}$	Yhat$_1$	Yhat$_2$
1	1	1	1
1	0	0	0
1	0	0	0
1	1	1	1
1	0	0	0
0	1	1	1
0	0	0	0
0	1	1	1
0	0	0	0
0	0	0	0

AUC (Yhat$_1$) = AUC (Yhat$_2$)

Yhat$_1$ vs. $\hat{y}^{(i)}$ = 1 ; Yhat$_2$ vs. $\hat{y}^{(i)}$ = 0.5



3. Random Model (AUC = 0.5)

(1) $y^{(i)} = 0$ → AUC = 0.2 — Sharp Point

(2) For this → (1) → 0, (0) → 1

mapped to

(Random) AUC = 1 - 0.2 = 0.8

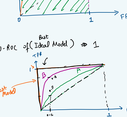
Ideas with ROC
highly imbalanced data → Anomaly search.

High Imbalance → F1, PR, F2

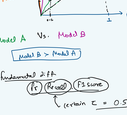
Precision Recall Curve (PRC)

AU-PRC

AUC-PRC measure reliable



KNN → Geometric Intuition



know → 1 nearest neighbour

(1) negative → tumor not present
(2) positive → tumor present

Tumor → Dog vs. Cat

K-NN

1. Supervised Algo.
2. Classification & Regression.
3. Instance based learning.

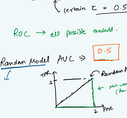
Algorithm.
Inputs: Domain, Distance, K, Training Data

$x^{(q)} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$

K = 5

(1) For each data point in Domain.
→ Compute dist $(x^{(q)}, x^{(i)})$
→ output x_i in a list.
(2) Sort the distances list in asc. order.
(3) pick K nearest neighbours from top.
(4) Majority Vote.

+ve → 3, -ve → 2 → Positive +ve



If Binary classification & if K = even = 6

+ve → 3, -ve → 3 → tie → random decision

Tie → wrong prediction

Robust even recommended in case of binary classification.

Disadvantage: Train time $O(1)$, test time → v.v.high.

(Test) → (1) Compute dist with n data. → $O(nd)$
(2) Sort distances → $O(n \log n)$
(3) Pick top K distances → $O(K)$
log(greatest number)

Test → $O(nd + n\log n)$

(for n Query) → $O(n(d \log n + 1))$
→ $O(n(d \log n))$

log(n) → 32 bits → 4 GB RAM

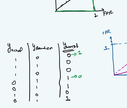
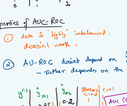
logistic regression test time complexity → $y_i = \frac{1}{1+e^{-(w^T x_i + b)}}$
$O(d)$

Hinge

n = 10
Test = 10
for i in range (1, n+1)
Test += i
Print (Test)

$1+2+3+\dots+n$ → $\frac{n(n+1)}{2}$ → $O(n^2)$
$O(n)$

Advantages:
→ Simplicity & Intuitive Algo.
→ Very Powerful
→ Non-linear Model.



Assumptions → Neighbourhood should share similar Label / homogeneous.
→ Similar Items