Topic 1

outlier removal methods

what is an outlier?

- A data point that is significantly different from other observations
- It is a point that doesn't belong with the main cluster of data

⇒ why care?

They can seriously skew your results.
An outlier can pull the mean (average) & warp the line a model tries to fit, leading to incorrect predictions

Method 1: 3 Sigma (3σ) Rule

- Based on the properties of a normal distribution (the BELL CURVE)

Key Assumption: your data should be roughly bell shaped for this work well

The Rule: Any data point that falls outside of 3 std. deviations from the mean is considered as outlier

— Working —

Calculate the mean (μ)
(Find the average of your data)

↓

Calculate Std. deviation (σ)

↓

Define boundaries

x = lower → $\mu - 3\sigma$ & upper → $\mu + 3\sigma$

↓

Any point $x <$ lower bound $x >$ upper bound is an outlier

---

pro → simple & fast, cons → not robust

Method 2: IQR method

A robust method not influenced by extreme values. It looks at the middle 50% of the data

Key Assumption: Doesn't require the data to be normally distributed

Rule: Any pt. outside the range defined by 1.5 x the IQR below Q_1 & above Q_3 → outlier

Sort → from smallest to largest

↓

Find Quartile

Q_1: The 25th percentile

Q_3: The 75th percentile

↓

calculate IQR (IQR = $Q_3 - Q_1$)

↓

Define boundary:

Lower = $Q_1 - 1.5 \times IQR$ & upper = $Q_3 + 1.5 \times IQR$

↓

Identify

Pro → very Robust to outliers

cons → slightly more steps to calculate

Topic - 2 : overfitting, underfitting & Generalization

The goal: A Generalized model

← The model learns the truth underlying pattern from the training data while ignoring the random noise

A generalized model performs well not only on the data it was trained on, but more importantly on new, unseen data

Analogy: A student who studies to truly understand the concepts (generalization) will ace both the practise test & the final exam

underfitting: (too simple model)

The model is not complex enough to capture underlying true trend in the data

This is a problem of high bias. The model makes overly simplistic assumptions about the data

The model has high error on Both Training & test data

overfitting - Too complex model

The model learns the training data too perfectly. It starts to memorize the random noise & outliers as if they were real patterns

This is a problem of high variance. The model is overly sensitive to the specific data it was trained on

The model has very low error on the training data high error on new data

The Bias Variance trade off

BIAS → Error from being too simple

Variance → Error from being too complex

The trade off → ~~As you ↓ bias~~

As you ↓ bias → may cause ↑ in variance

goal is to find the perfect balance

TOPIC-3 SVM - Alg. used for classification

Maximum margin classifier

main job to find the best possible boundary that seperates 2 classes

Key idea: It doesn't just find any line that seperates the data it finds the one that creates the widest possible empty space b/w 2 classes. This is called maximizing the margin

Key terminology

Hyperplane: The decision boundary itself

└ In 2D → It is a plane

3D → it is a plane

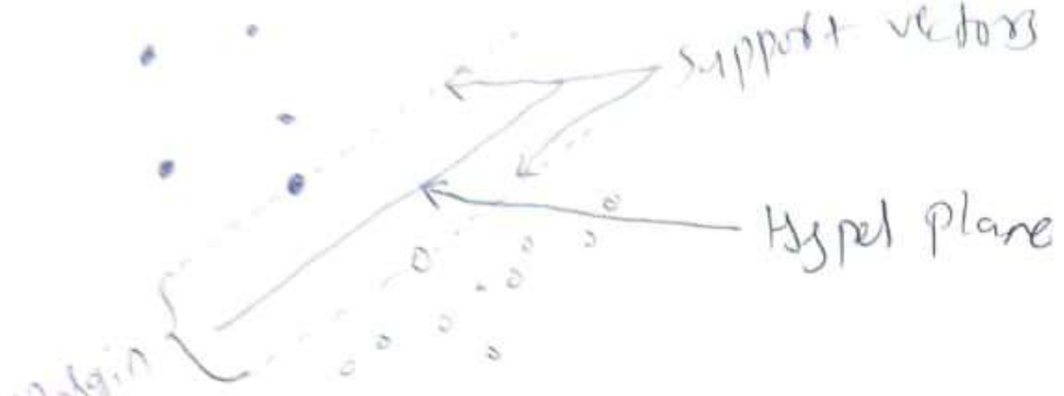
Support vectors:

data points closest to hyperplane

They lie on the edge of margin

Margin: This is the empty space

└ goal of SVM to make empty space as wide as



working

plot the data points

↓

Find a separating hyperplane

↓

IDentify support vectors & margin

↓

Maximize the margin

↓

classify new data

Margin

Hold margin

- very strict, data perfectly separatable

- Mistakes allowed = 0
- extremely sensitive to outliers

Soft Margin

much more flexible

The trade off

Using the 'C' Hyper parameter

High 'C' → Acts more like strict Hold margin

Low 'C' - more tolerant

TOPIC - 4

Hyperparameter tuning

What is Hyperparameter tuning?

A setting or configuration for a model that is chosen before the training process begins

The model doesnot learn hyperparameters the data scientist set it

Eg:-

The C value in SVM

no. of trees in Random Forest

learning rate for a neural network

Why tuning is necessary?

The performance of a model is extremely sensitive to its Hyperparameter settings.

To find the specific combination of hyperparameters → to find generalized model

Common tuning techniques

Grid search

Create a grid of possible values you want to test for each hyperparameter

Grid search exhaustively trains & evaluates a model for every single possible combination

pros → you are guaranteed to find best combination

cons → slow & expensive

Randomized Search

It tests a fixed no. of random combinations from the range of values you provide

pros → faster & more efficient

cons → not guaranteed to find the absolute best combination

Grid Search Step

→ Define the hyperparameter grid
→ Create all combination
→ Evaluate each combination
→ Select the best combination

Randomized Search

→ Define hyperparameter space
→ Set the no. of iterations
→ Randomly sample & evaluate
→ Select the best combination

Topic-5 Cross validation

Get a reliable & stable estimate of a model performance

It prevents data leakage

The k-fold cross validation process

The initial split

The very first step is to partition your entire dataset into a training set & a test set

The test set is locked away. All the following steps happen only on the training set

2 Partition into 'k' folds

The training set is divided into 'k' equal sized, non overlapping subsets called 'folds' (k is usually 5 or 10)

3) The iteration loop

The process iterates 'k' times, in each round
- A single unique fold chosen as the validation set
- The remaining k-1 folds used as the training data
- The model is trained on the training data & then scored on the validation fold

4) Aggregate the scores

After 'k' rounds you will have 'k' performance scores

The final cross validation score is the average of these k scores