Feature Selection

Why we need feature selection

- Curse of dimensionality
  - If we have certain no. and getting optimal soln we increase feature. After optimal # model of performance ↓
    ⇓
    why?
    ⇓
    Sparsity ⇒ on increasing dimension distance betn two data points increases sparsity
- Computation complexity
- Interpretability
  - difficult to inference of ML model.

Techniques (Types) of FC

- Filter based Tech.
  - Variance threshold — constant feat / quasi const.
  - Correlation based
  - ANOVA
  - chisq
  - mutual info.
- Wrapper methods
  - Exhaustive FS
  - Forward selection
  - Backword elimination
  - Recursive feature elimination
- Embeded methods
  - Lasso, Ridge, Elastic → for linear data
  - DT, RF, GB --- → for non linear data.
- Hybrid technique

If comes in Hybrid

Filter based feature selection

Methods that use statistical to score each feature independently, then select a subset of feature based on their scores.

Called as "Filter" methods because they essentially filter out the features that do not meet some criterion.

Step1 →

Deleting duplicate features
- exactly same column with value

~~1) Varian~~

1) Variance Threshold

- Constant : Variance = 0
  Ex:- B: 1, 1, 1, ---
- quasi's constant
  variance → 0
  Ex:- B: 1, 1 ... 999, 5 5 5 ... 5
  999 are 1   5 are 5

We will set-up a variance threshold (ex:- 0.05) if variance of each columns < var_th which will dropped

→ before applying this we must apply normalize or standarize their values.
⇓
If data is standarized and normalize threshold should be in [0.1, 0.01]

When we should not apply variance threshold blindly?

1. Ignores Target Variables:-
∵ univariate method, evaluates features independently and doesn't consider the relationship b/w each feature ~~that a high variance~~ and target variable.
↓
It may keep irrelevant feature that have high variance but not relationship with target or vice-versa.

2. Ignores feature interactions:-
A feature with low variance become very important (informative) when combined with another feature.

3. Sensitive to data scaling:-

4. Selecting threshold value is big challenge.

3. Correlation

Find correlation bet^n each pair of independent variable & we will discard features based on corr. coef value is ~~based~~ ± threshold (Ex: ±0.8)

Disadvantages of correlation

(1) Linear Assumption
↳ It does not capture non linear behaviour while removing, that can be misleading.

(2) Doesn't capture complex relationship:
↳ may not capture complex relationship involving more than two variables.

(3) Threshold Determination.
↳ Depending on problem & dataset.

(4) Sensitive outliers:-

ANOVA.

→ It applies only when all independent features → numerical
dependent features → catego > 2 categories

→ to apply ANOVA
↳ we have to calculate F-ratio
↳ MS between / MS within.

MS between = SS between / (K-1) — # categories in dependent
MS within = SS within / (n-K) — #rows, # categories in dep. var.

Col-1	y
1	S
2	L
3	W
4	S
7	W

→

S	L	W
1	2	3
4	.	7

↑ unstacking.

grand mean ($\bar{\bar{x}}$) = mean of this entire col.

$\bar{x}_S$ = mean of S class after unstacking

$\bar{x}_L$ =

$\bar{x}_W$ =

$SS_{within} = (1-\bar{x}_S)^2 + (4-\bar{x}_S)^2$ — Subtracting mean of each class from each observation of belonging to that class

$+ (2-\bar{x}_L)^2$

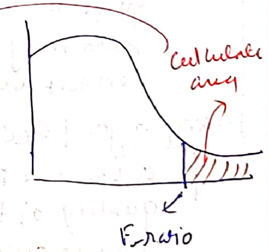
$+ (3-\bar{x}_W)^2 + (7-\bar{x}_W)^2$

we do for each class.

$SS_{between} = n_1(\bar{x}_S - \bar{\bar{x}})^2 + n_2(\bar{x}_L - \bar{\bar{x}})^2 + n_3(\bar{x}_W - \bar{\bar{x}})^2$

- $\bar{x}_S$: category mean for that class
- $\bar{\bar{x}}$: grand mean (mean of column)
- n: # observation under particular class..

$$\frac{SSW/(K-1)}{SSB/(n-K)} = \text{F-ratio or F-stat} \rightarrow \text{It follows f-dist.}$$



< 0.05
⇓
there is relationship b/w F → Y.

Chi sq

It is used to find selected ~~best~~ columns based on categorical vs categorical
(categorical → i/p cols, categorical → o/p cols)

Ex: $f_1 \rightarrow Y$.

Ex:- is there relationship b/w sex → survived?

⇓

Step 1 (contingency table)

	0	1
M	468	109
F	81	283

→ observed value.

⇓

Step 2 (expected value)

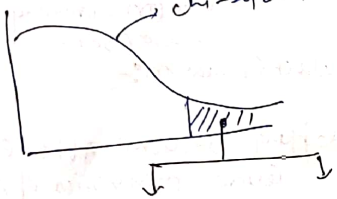
	0	1
M	193	355
F	120	221

$\frac{(81+468)*(81+283)}{\text{Total Sum.}}$ ⇒ similarly we calculate for all cell.

Now we have to find diff b/w (observed value) vs (expected value)

⇓

$$\text{Chi-sq-Stats} = \sum_{i=1} \frac{(\text{observed val}_i - \text{Expected val}_i)^2}{(\text{Expected val})_i}$$



P < 0.05
⇓
reject H0
⇓
there is relationship b/w ...

Disadvantage

Categorical Data only

- Can only be used with cate. data
- not suitable for cont. data unless they have been discretized into categories

independence of observations

- Chi-sq test assumes that the observations are independent of each other.
- Thi

Disadvantage of Filter based Selection

- we are only selecting features based on $x_1 \rightarrow y$ relationship but we are not finding based on $x_1 x_2 \rightarrow y$ or other combinations of features relationship with y

⇓ solution

wrapper method

It involve using a predictive model to score the combination of features.

steps involved

- subset generation → with several technique Adding one by one (or) removing one by one or so on.
- Subset Evaluation:-
- stopping criterion:- After evaluating certain possibilities of subset no further improvement than we will stop the process.

wrapper method

- Exhaustive FS
- Forward selection
- backward elimination
- Recursive feature elimination

Exhaustive FS / best subset selection

→ trying all the subsets along with Model development ~~that subset~~ that subset whose accuracy is best, that subset will be selected in FS.

Disadvantage

→ n cols → need $(2^n - 1)$ models to be train.

- Computational Complexity
- Risk of over fitting
  - Feature combination that performs best on the training data may not perform best on test data (unseen data)
- Req. a good Evaluation metric.
  - Exhaustive FS depends on the quality of the evaluation metric used to assess goodness of a feature subset.

*** what is cross validation (cv=5)?**

we split our data 5 times in train & test data set. we find accuracy for each split and at the end we will find avg score.

sequential Backward Selection / Elimination

F_1 F_2 F_3 F_4 → model

⇓

we develop 4 model by removing ~~[illegible]~~ each in each combination

⇓

Select Model having 3 features having highest score.

⇓

3 model by removing each in each combination.

⇓

~~select Model having~~

F_1 F_2 F_3 F_4 → model → 0.89 All

⇓

f_1 F_2 F_3 F_4 → model → 0.81

F_1 f_2 F_3 F_4 → " " 0.71

F_1 F_2 f_3 F_4 → model → 0.91 (remove this feature)

F_1 F_2 F_3 f_4 → Model → 0.65

f_1 F_2 F_4 → model → 0.79

F_1 f_2 F_4 → model → 0.81

F_1 F_2 f_4 → model → 0.83 (remove)

f_1 F_2 → model → 0.63

F_1 f_2 → model → 0.53

remove.

⇓

Step 2

Now we will select best model

$f_1 f_2 f_3 f_4$	0.89
$f_1 f_2 f_4$	0.91
$f_1 f_2$	0.83
f_2	0.63

this is the best ($f_1 f_2 f_4$)

Advantage

- faster
  - ∵ for n columns we require $\frac{n(n+1)}{2}$ models to check
- We may miss best combination because After removing any feature, that feature can't be used in next combing iteration due to which it may loose best combination

Sequential forward selection.

Here we add in each iteration

Let's we have

F_1 F_2 F_3 F_4

Add iteration 1

f_1 → 0.63 | f_2 → 0.51 | f_3 → 0.43 | f_4 → 0.49

Add iteration 2

$f_1 f_2$ → 0.63 | $f_1 f_3$ → 0.71 | $f_1 f_4$ → 0.80 ✓

Add in iteration 3.

$f_1 f_4 f_2$ → 0.81 | $f_1 f_4 f_3$ → 0.85 ✓

And in iteration 4

$f_1 f_4 f_3 f_2$ → 0.83

Now, select all iteration's best

f_1 → 0.63

$f_1 f_4$ → 0.80

$f_1 f_4 f_3$ → 0.85 ✓ best, final.

$f_1 f_2 f_3 f_4$ → 0.83

Advantage

- faster vs Exhaustive
  - $\frac{n(n+1)}{2}$ are models need to be train.

Ex hausting → $2^n - 1$ models

backword → $\frac{n(n+1)}{2}$

forward → $\frac{n(n+1)}{2}$

Disadvantage

→ Here also we may skip best possible combination.

Ex:- f_1 got selected in 1st iteration & it's getting clubbed f_2, f_3, f_4 then. (f_2, f_4) Combination we are not cheking.

Feature Selection using Embedded.

Disadvantage of

Filter:- Missing mutual feature interaction while selecting features based on perform.

Wrapper:- slow & missing some best combination.

↓↓ solve.

Embeded

works as

while training only it gives feature importance to train that model. due to which its faster as well as taking consideration of all possible combination.

ML Algo

If any ML algo have below attribute we can use as embeded method.

- Coef_
  - Linear_reg.
  - Logistic reg.
  - LASSO
  - Ridge
  - ElasticNet

  ↓↓ Works well for linear data

- feature_imp_
  - DT
  - RF
  - gradient boosting.

  ↓↓ Works well for Non-linear data.

Ex:-

$LPA = \beta_0 + (\beta_1) cgpa + (\beta_2) iq$

then Co-eff only tells that by increasing 1 unit respective factor, How many units target variable will increase

So, somewhat Co-eff- of each feature tells their importance

↓↓

But it give true picture of Feature importance of all assumption of suits very well.

- linearity
- independence
- Homoscedality
- Normality
- No Multicollinearity.

FS on Regularized Model

Regularized linear model are linear Models that include a penalty term in the loss function during training.

The penalty term discourage the learning of a too complex model, which can help to prevent Overfitting

- Ridge
- Lasso
  - best to use for FS.
  - ∵ It keeps imp. feature's Co-eff as "Non-zero" when as make co-eff. of other feature as "zero"
- ElasticNet

→ If in all three algo there is an attribute called feature_importance

→ that will give importance of each feature under that Model.

Wrapper method's
Recursive Feature Elimination

its kind of hybrid technique.

Wrapper embeded

$f_1\ f_2\ f_3\ f_4$ → Model → remove it based on weak feature importance

$f_1\ f_2\ f_3$ → Model → x → $f_1\ f_3$ → Model → feature_imp

x

∴ this way it is best feature.

we can do it this same process to apply on rest feature to find 2nd best feature.

~~Mutual info~~
&
~~filter method~~

Advantages of Embeded techniq

→ Performance:- they are generally more accurate than filter methods since they take the interactions b/w features into account.

→ Efficiency:- ↑ than wrapper ∵ they fit model only one

→ less prone to Overfitting:- due to regularization, less prone to overfitting.

Disadvantage of Embeded

→ Model specific

→ Complexity

→ Tuning Required : value of λ in Lasso & ridge

→ Stability

Mutual information in Filter Method.

MI is measure of dependency b/w 2 variables

It is fundamental quantity in information theory

$$MI = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \left[\frac{P(x,y)}{P(x)\,P(y)} \right]$$

where,

$P(x,y)$ → joint Prob of x & y

$P(x)$ → marginal Prob of x

$P(y)$ → marginal Prob of y

$P(x, y=0)$ → Probability of $y=0$ being $x=M$ or other diff cases based on value of x & y

X	y
M	0
F	1
M	0
F	0
M	1

⇒

(cross table)

	survived = 0	survived = 1	
M	2/5	1/5	3/5
F	1/5	1/5	2/5
	3/5	2/5	

$P(x=M, y=0) = 2/5$

Joint probability for this condition..

these are individual or Marginal probability

to find MI sex → we calculate one for each cells.

$$\left[\frac{2}{5}\log\left(\frac{2/5}{3/5 \times 2/5}\right)\right] + \frac{1}{5}\log\left(\frac{1/5}{3/5 \times 2/5}\right) + \frac{1}{5}\log\left(\frac{1/5}{3/5 \times 2/5}\right) + \frac{1}{5}\log\left(\frac{1/5}{3/5 \times 2/5}\right)$$

If MI (sex) ↑ then it higher important feature

Few points about MI

→ It is Non-negative;- always zero or +ve with zero indicating that the variables are independent

→ It is symmetric:
$MI(X,Y) = MI(Y,X)$

→ It can capture any kind of statistical dependency:
Unlike correlation, which only can capture linear relationship. It can capture any kind of relationship, include non linear ones.

How to deal with Numerical Variables?

↳ unlike chi-sq. it is able to work on numerical data.

Ex:-

age	survived
35	0
60	1
52	1
16	1
27	0

Internally it will convert into discrete or bins.

Disadvantage

→ 1. Estimation Difficulty:- When the dimensionality of the data is high or the number of sample is low

→ 2. Assumes Large Sample Size MI works best with large sample size

→ 3. Computationally Intensive:-

→ 4. Difficulty with Continous Variables

→ 5. No direct indication of the Nature of Relationships

→ 6. Doesn't Account for Redundancy