Lecture-39

Missing values, imputation & Surrogate result.

Imputation:

Mean, Regression. [predict missing points]

└ Full information imputation.

Using all other feature to impute values.

Multiple imputation:- use distribution to predict values.

↓

Follows 3 steps.

1. similar to single Imputation, missing values are imputed. However, the imputed values are drawn m times from a distribution rather than just once. → gives (m - datasets)
2. Analysis: At end of this step, there should be m - analyses.
3. pooling: m results are consolidates into one result by calculating mean, variance, confidence intervals of the variable.

↓

~~by~~ i single Imputation does not take care of uncertainity in imputed values whereas multiple imputation accounts for the uncertainity and range of values that the true values could have been.

→ Fitting it into a distribution and then sampling values using mean and variance of distribution

↑imputation. # Lecture-39: Surrogate splits

- New categorical ~~feature~~ value "missing".

Age		Age
20	⇒	20
30		30
..		"missing"
..		"missing"
40		40

} There are some reasons, why there is missing values in our data. If we impute it, we are going to add noise in the feature. So, It's best to just ~~leave~~ replace those values by "missing".

Surrogate splits

In decision tree, when a split is made, It depends on one variable. but what if that variable is missing. Suppose, we want to predict whethers customers will be kept or lost next year.

order < 6.6

Y (Lose) N (Keep)

→ Surrogate split will try to predict actual split:

→ Another decision tree will be created to predict your split.

order ≤ 6.5 | orders ≥ 6.5 → Fake decision tree

- Fragment (specific to trees)

$X_3 < 5$?