Train In Data

Anova

Anova

Anova tests the hypothesis that 2 or more samples have the same mean.

- Samples are independent
- Samples are normally distributed
- Homogeneity of variance

One way Anova: classification

Var1 (y=0)	Var1 (y=1)
Obs 1	Obs 5
Obs 2	Obs 6
Obs 3	Obs 7
…	…
Mean 0	Mean 1
Variance 0	Variance 1

One way Anova: classification

Var1 (y=0)	Var1 (y=1)
Obs 1	Obs 5
Obs 2	Obs 6
Obs 3	Obs 7
…	…
Mean 0	Mean 1
Variance 0	Variance 1

One way Anova: classification

Var1 (y=0)	Var1 (y=1)
Obs 1	Obs 5
Obs 2	Obs 6
Obs 3	Obs 7
…	…
Mean 0	Mean 1
Variance 0	Variance 1

Great mean

One way Anova: classification

Var1 (y=0)	Var1 (y=1)
Obs 1	Obs 5
Obs 2	Obs 6
Obs 3	Obs 7
…	…
Mean 0	Mean 1
Variance 0	Variance 1

Great mean

- **Model Sum of Squares**:
  - $\sum Obs(Mean - great\ mean)^2$
- Or
  - $\#\ Obs(0)\ \times\ (mean0 - great\ mean)^2 +$
    $\#\ Obs(1)\ \times (mean1 - great\ mean)^2$
- **Residual sum of squares**:
  - $\sum(Obs(0) - mean0)^2 + \sum(Obs(1) - mean1)^2$

One way Anova: classification

Var1 (y=0)	Var1 (y=1)
Obs 1	Obs 5
Obs 2	Obs 6
Obs 3	Obs 7
…	…
Mean 0	Mean 1
Variance 0	Variance 1

Great mean

- Mean Squares model: $\frac{(\textbf{Model Sum of Squares})}{(\text{Columns } -1)}$
- Mean Squares error: $\frac{(\textbf{Residual sum of squares})}{(\text{Total obs } -1)}$
- F-statistic: $\frac{(\text{Mean Squares model})}{(\text{Mean Squares error})}$

One way Anova: Regression

Var1	Target
Obs 1	Value 1
Obs 2	Value 2
Obs 3	Value 3
…	…

- Correlation coefficient between variable and target.
- Convert the correlation into a p-value

- https://github.com/scikit-learn/scikit-learn/blob/0fb307bf3/sklearn/feature_selection/_univariate_selection.py#L232

Anova: Scikit-learn

- **F_classif or f_regression**: rank features ➔ smallest the p-value biggest importance

- **SelectKBest**: select best k features
- **SelectPercentile**: select features in top percentile

Train In Data

THANK YOU

www.trainindata.com