Practical Statistics for Machine Learning.

Data Types

1 Continous : Data that can take on any interval.

2 Discrete : Data that can take only integer values

3 Categorical : Data that can take only a specific set of values representing a set of possible categories

4 Binary : Data with only two possible categories (0 / 1, true / false).

5 Ordinal : Categorical data that has explicit ordering.

Examples :

1 Continous : Wind speed
Time duration

2 Discrete : Count of occurrence of event
Number of persons in a population

3 Categorical : List of states in country

4 Binary : Spam mails

5 Ordinal : T-Shirt Sizes (S, M, L).

Structured Data

- Numerical
  - Continous
  - Discrete
- Categorical
  - Ordinal
  - Binary.

Why do we want to classify data?

1. Storage and indexing can be optimized.
2. The possible values a categorical variable can take are enforced in a software (e.g. enum)
3. Knowing the nature of data can help us in plotting a chart or fitting a model.

Estimates for Location.

1. Mean: The sum of all values divided by the number of values.
2. Weighted Mean: The sum of all values times the weight of each observation as sum
3. Median: The middle value of the data
4. Weighted Median: The value such that one half of the sum of weights lies above & below the sorted data.
5. Trimmed Mean: The average of all values after dropping a fixed number of extreme values

6 Robust : Not sensitive to extreme values

7 Outlier : A datapoint which is very different from other data points.

Metrics : The estimated values calculated from the data to draw a distinction from the data to other means or features within data.

- Mean = $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$.

N = Total no. of records or <u>Population</u>.

n = Total no. of records in <u>Sample</u>

$n \in N$

n is a subset of N.

Mean is known to be prone to extreme values or outliers.

∴ We use a variation called <u>Trimmed Mean</u>

$$\text{Trimmed Mean} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_i}{n-2p}$$

It eliminates extreme values p on both sides yielding a very robust metric.

Scanned with CamScanner

Weighted Mean = $\bar{x}_w = \dfrac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

1. We use weighted mean because some variables are more intrinsic than others and highly variable values are given lower weights than others

2. The data does not represent the groups we want to measure equally.

Median: Sort the values in ascending order
If n is odd then find middle term
If n is even then find average of the middle terms.

Median is a robust estimate because it is not affected by outliers

Estimates of Variability

1. Deviations: The difference between observed values and estimate of location

2. Variance: The sum of squared deviations from mean divided by n-1 where n is the number of data points.

3. Standard deviation: Measure of dispersion within data ie square root of variance

4. Range: The difference between the largest and smallest value in the data

5 Percentile : The value of P percent of the values take on this value or less and (100 - P) percent take on this value or more.

6 I- Quantile Range : The difference between 75th percentile (3-IQR) and 25th percentile (1-IQR)

Mean Absolute deviation = $\frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$

Standard deviation = $\sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$

$S = \sqrt{\text{variance}}$

Degree of Freedom :

1 We use n-1 and not n terms in calculating variance.

2 If we use intuitive denominator n in variance, we end up with a biased estimate. However, if we divide by n-1 instead of n, the standard deviation becomes an unbiased estimate.

3 The reason for biased estimate with denominator n is that formula for standard deviation is having mean of n terms.

A robust estimate of variability is the median absolute deviation (MAD).

$MAD = \text{Median}(|x_1 - m|, |x_2 - m| \ldots)$
where m is the median

$\therefore \sigma > \text{Mean AD}$.

Estimates Based on Percentiles

When the data is sorted the estimates are called order statistics

Range = Largest - Smallest.

The p^{th} percentile is a value such that atleast p-percent of the values take on this value or less and atleast (100-P) percent values is more than that.

Inter-quartile range: The difference between the 75^{th} percentile and 25^{th} percentile

Example: 3, 1, 5, 3, 6, 7, 2, 9
=> 1, 2, 3, 3, 5, 6, 7, 9

$\frac{2+3}{2} = 2.5$ $\frac{6+7}{2} = 6.5$

$\therefore IQR = 6.5 - 2.5 = 4.0$

$\text{Percentile}(P) = (1-w)\, x_{(j)} + w\, x_{(j+1)}$

$100 * \frac{j}{n} \leq P < 100 * \left(\frac{j+1}{n}\right)$.