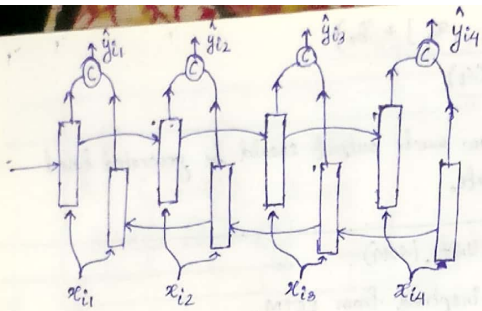


→ Time.

Ⓒ - concatenate

Bidirectional RNN

Exploratory Data Analysis (EDA).

Pair-plots

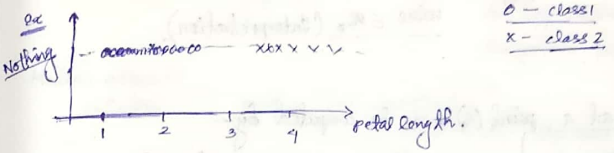
* When we have many features of the datapoint & we want to visualize it, since we humans can't see beyond 3-D, we do a small hack.
* We take every possible pair of features & plot the pts. Visualising and analysing can give a sense of how the data is distributed in higher dimensions.
* Can be used to decide which pair of feature best separate the datapoints.

Limitations of pair-plot

* If the dimensionality of data is small, pair-plots do a pretty decent job & fairly easy to understand.
  But higher dimensionality data would have lots of pair-plots as total - $\binom{n}{2}$. Hence we'll be having a hard time analysing them all.
* Also, pair-plots won't help if there is some pattern that's formed only in higher dimensions.

Histogram & PDF

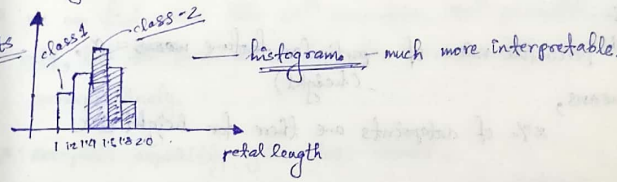
* 1-D scatter plots by themselves are hard to interpret.



* We can't really say how many points are there as many are overlapping.
* Also, 1-class point may overlap another & we may have no idea about that.

So, in-order to make things more interpretable,

* we break the x-axis into smaller buckets.
  say for ex, we divide 1 ——— 2 into 5 buckets.
  $\{[1-1.2], [1.2-1.4], [1.4-1.6], [1.6-1.8], [1.8-2.0]\}$.
* say 4 pts b/w 1 & 1.2, 3 pts b/w 1.2 - 1.4. & so on.
* In graph we represent the number of point along y-axis.



histograms — much more interpretable.

* A smooth approximation of the histograms is the probability density function (pdf).
* PDF obtained by KDE (kernel density estimation) of histograms.
* Area under PDF = 1.
* pdf at a point can be thought of the %age of total data points at that value.

(Or)

* probability of finding a datapoint at that value.

CDF (Cumulative density function).

* CDF value at a point (x_0) = %age of datapoints that have value $\leq x_0$. (Interpretation).

* CDF value at a point (x_0) can be computed by—

$$\int_{-\infty}^{x_0} pdf \, dx$$

i.e. area under the curve of pdf till x_0.

* CDF can be very handy in coming up with threshold values for simple if-else models as it is very easy to retrieve values, less than which all the datapoints of a class lie.

---

* Mean & standard deviation are prone to outliers.
* Median gets corrupted only if more than 50% of the points are corrupted.

Percentiles—

x^{th} percentile value of a particular feature (height) = α.

That means, x% of datapoints are there for height $\leq \alpha$.

* So, if want say 80th percentile value of a feature. Then we can simply refer to the CDF plot & find the value of that feature corresponding to which the CDF is 0.8.
* We can try this on the whole dataset or separately on different classes.
* Doing separately will give more insight on each individual class's distributions which helps in classification. — which is the Objective

Median Absolute deviation

In std-dev, $\sigma = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{1/2}$ - Its kind of calculating the avg. distance of the datapoints from the mean.

Median absolute deviation = $\text{median}\left(|x_i - \text{median}|\right)_{i=1}^{n}$

($|x_i - \text{median}|$: absolute deviation)

Its kind of similar meaning like above but we're using median instead of mean. Hence, more robust to outliers.

Inter-Quartile Range (IQR)

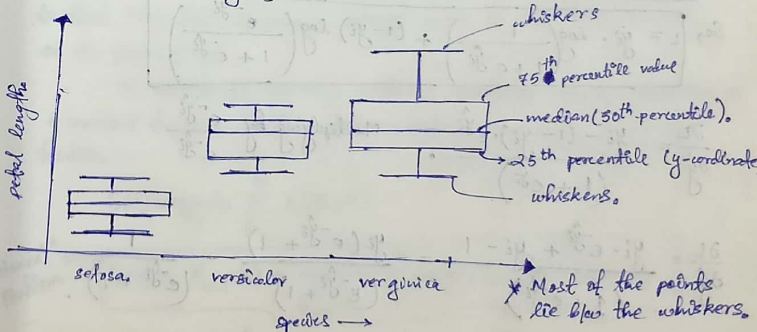
It can be calculated by

$(75^{th}$ percentile $- 25^{th}$ percentile$)$.

Gives a range where the central 50% points lie.

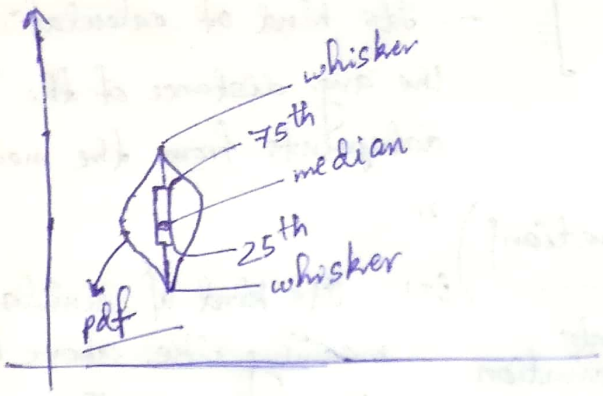
Box-plots

* We can find values like 25^{th} percentile, 75^{th} percentile using cdf, but for that we'll have to draw parallel lines correspondingly.
* Box-plots explicitly give these values.



* Most of the points lie b/w the whiskers.

Violin-plots → (box plot + pdf)



* Always perform analysis oriented towards solving the obje
Don't deviate.

ŷi
an

80,