PRML Assignment 1 Report

March 19th, 2019

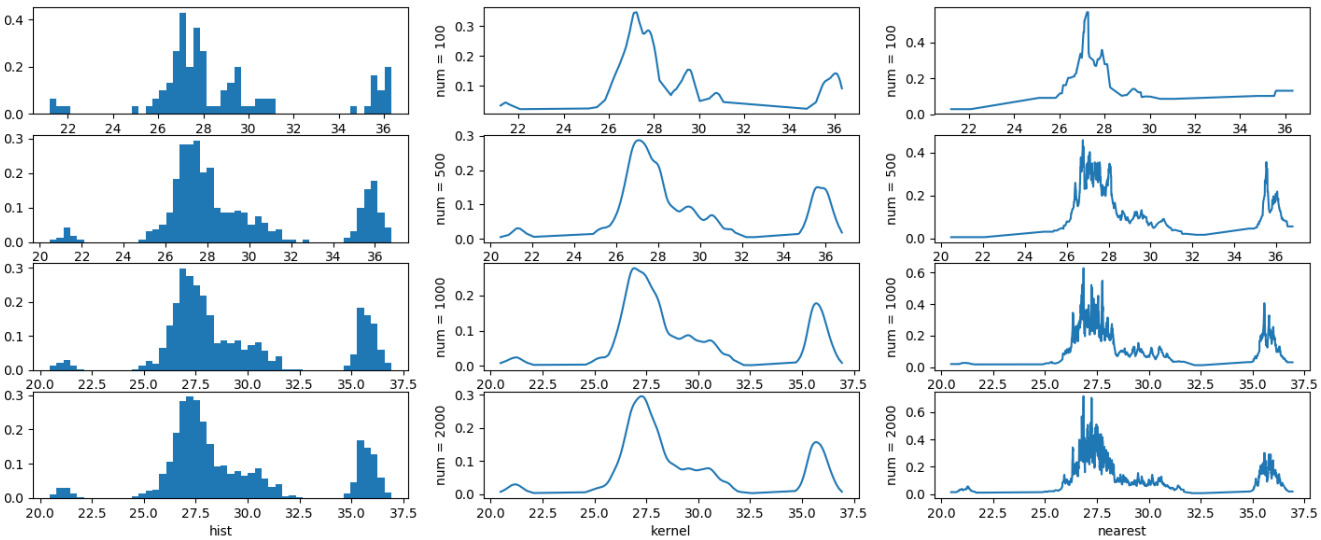
Part 1

How the number of data influence the quality of the estimation?

The following plot shows the estimation of the three given method using different data numbers.

We can easily figure out from the plot that as the number is increasing **`histogram estimation` and `kernel density estimation` work better and the curves become smoother**.

On the other hand, **`nearest neighbor estimation` appear to be worse and the curves become very spiky**.

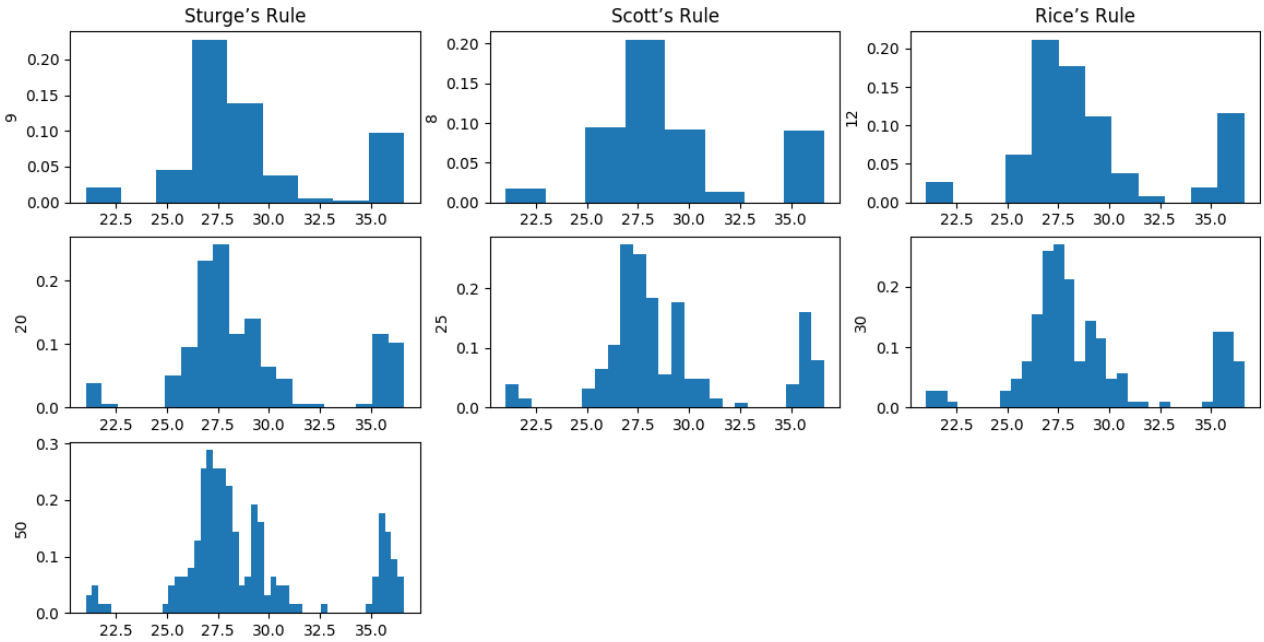


Part 2

How could you pick the best (or good) choice for this number of bins?

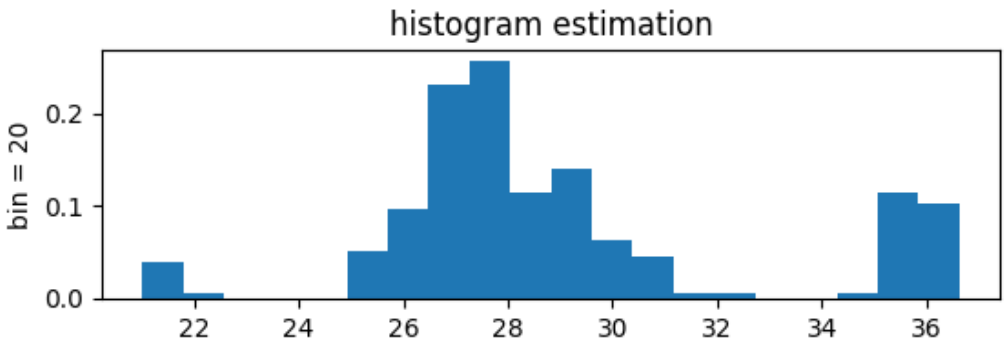
As far as I see, the best method to pick bin number is to **judge by eyes**.
However, there are still many formulas for picking bin number, so I tried three of them and pick out the best by eyes.

`Sturge's Rule` and `Scott's Rule` lead to too much loss in shape, **but `Rice's Rule` seems fine**.



Draw the best histogram (num_data=200).

I picked the bin number based on `Rice's Rule` and did some adjustment by hand.



Part 3

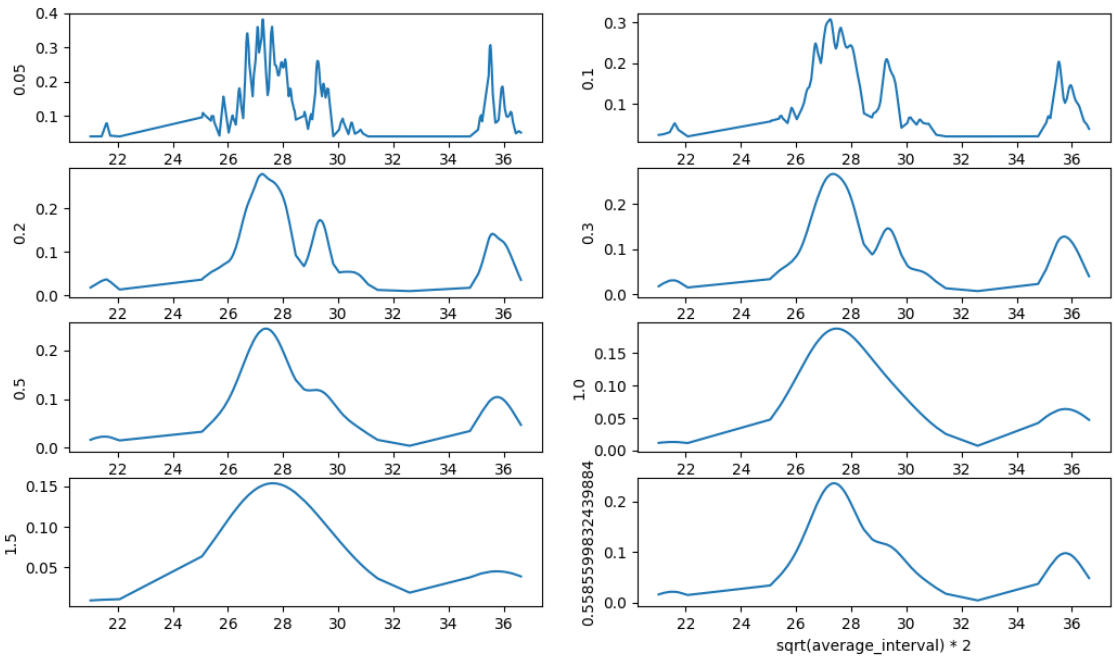
Answer if you have a clue of how to choose `h` .

According to the features of `Gaussian kernel` , we can know that the smaller the `h` is, the more `p(x)` is influenced by its nearby data than other data.

Therefore **I invent a formula to pick `h`**, based on the average interval of data.

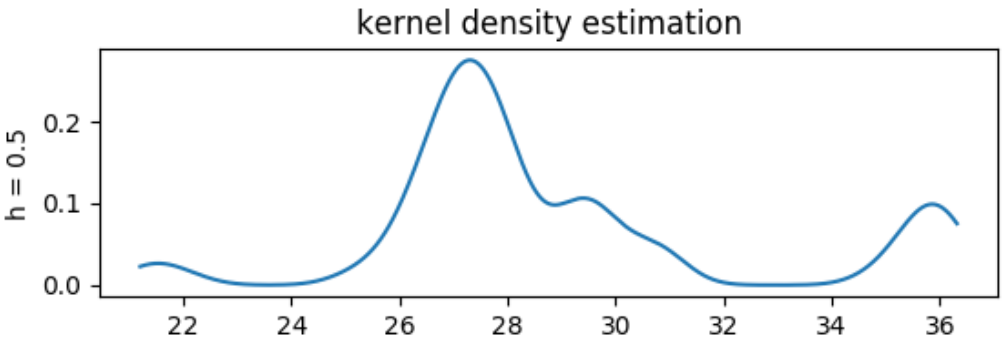
```
h = 2 * sqrt( sum( data[i] - data[i-1] )/( N-1 ) )
```

This formula works out fine no matter how large or how small the data set is.



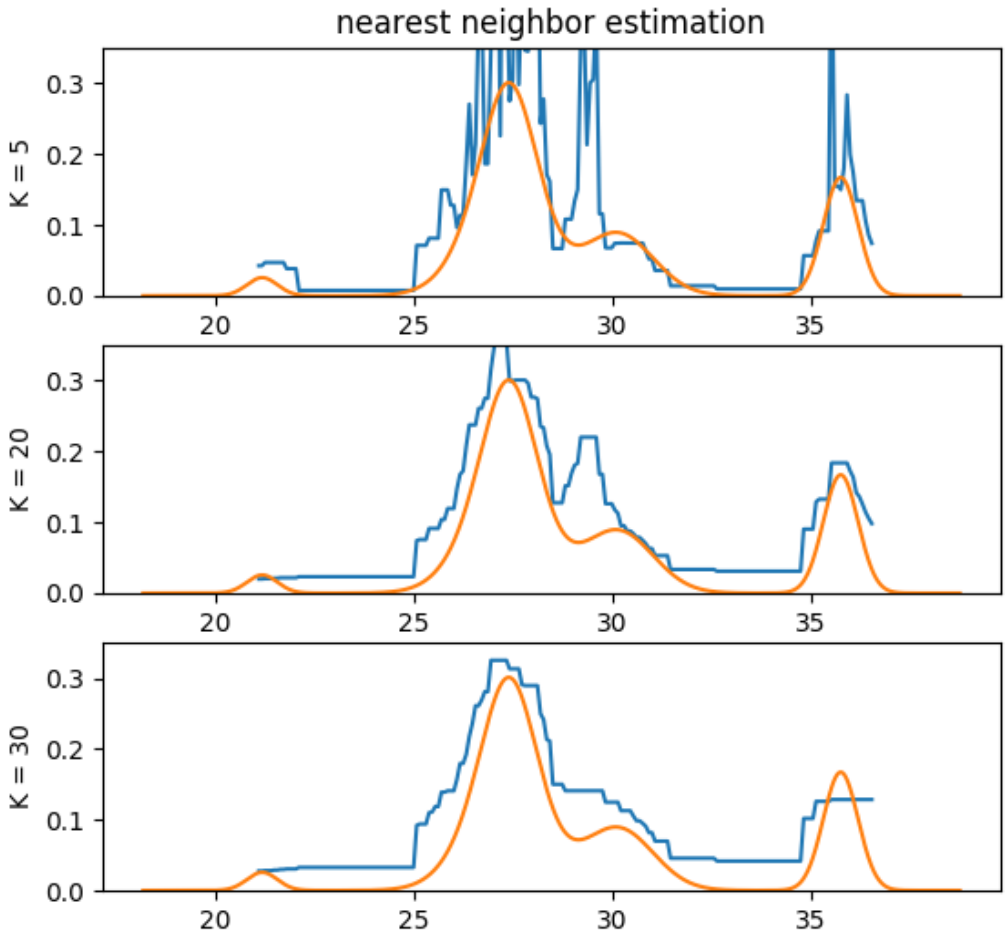
Plot the best estimate you could achieve with num_data=100.

I picked `h` based on my formula and did some adjustment by hand.



Part 4

Plot an illustration as Figure 2.26 in the text book (num_data=200).



Please show that the nearest neighbor method does not always yield a valid distribution.

Obviously, when `K` is small, `p(x)` will become too large if the data density is high, thus making the integral over all space overflow.
For intance, the second plot from the picture above shows that estimated `p(x)` is larger than true `p(x)` almost in all space, so it won't be a valid distribution.