1 Preprocessing

Given the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

- Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.
- How might you determine outliers in the data?
- What other methods are there for data smoothing?

2 Binning

Suppose a group of 12 sales price records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215. Partition them into three bins by each of the following methods:

- equal-frequency (equal-depth) partitioning
- equal-width partitioning

3 Relief

Given the data set X with three input features and one output feature representing the classification of samples:

I1	I2	I3	O
2.5	1.6	5.9	0
7.2	4.3	2.1	1
3.4	5.8	1.6	1
5.6	3.6	6.8	0
4.8	7.2	3.1	1
8.1	4.9	8.3	0
6.3	4.8	2.4	1

- Rank the features using a comparison of means and variances.
- Rank the features using Relief algorithm. Use all samples for the algorithm (m = 7).

4 PCA

Given 3 data points in 2-d space, (1, 1), (2, 2) and (3, 3),

- What is the first principle component?
- If we want to project the original data points into 1-d space by principle component you choose, what is the variance of the projected data?

5 Gradient descent (extra for experts)

Suppose we want to train a linear model $f(x) = ax + b$. We will use gradient descent to minimise the sum of squared errors over a training set $X = \{(x_i, y_i)\}_{i=1}^{N}$ of size N:

$$L(a, b; X) = \frac{1}{2} \sum_{i=1}^{N} (y_i - (ax_i + b))^2$$

Calculate the partial derivatives of the loss L with respect to a and b. In other words, calculate $\frac{\partial L}{\partial a}$ and $\frac{\partial L}{\partial b}$.