Retake Pattern Recognition
Wednesday, April 20, 2022
17.00-20.00 hours

Question 1: Mixed Questions (20 points)

(a) The simplest model that fits the description is:

$$\Delta \text{ chol} = a \times \text{dosage} + b \times \text{gender} \times \text{dosage}$$

(b)
1. Assign each point to the nearest cluster center.
2. Compute the cluster centers as the mean of all points that have been assigned to that cluster.

The algorithm stops if no reassignments have taken place in two consecutive iterations.

(c) Approaches that work well in low dimensions break down in high dimensional space. An example is estimating the joint probability distribution of a number of discrete random variables. The saturated model works fine with 2 or 3 variables, but not so much if you have, say, 20 variables. If the variables are all binary, there are still 2^{20} is more than one million possible value assignments, and hence more than one million probabilities to estimate. Making independence assumption could help in this case.

(d) The SVM prefers the line that maximizes the distance to the point closest to it. The intuition is that points close to the line represent cases about which we are uncertain, and we want to avoid such uncertain cases as much as possible.

Question 2: Linear Regression (20 points)

(a)

$$\text{SSE} = \sum_{n=1}^{50} (\text{pr}_n - a - b\ \text{dm}_n)^2$$

(b) If the grade for data mining increases with one point, then the expected grade for pattern recognition increases with 0.66 points.

(c) 61%, because $R^2 = 0.61$.

(d)

$$3.7 + 5 \times 0.66 = 7$$

(e) The regression line goes through the point of means. Hence

$$\begin{aligned} 7.73 &= 3.7 + 0.66x \\ 0.66x &= 7.73 - 3.7 \\ 0.66x &= 4.03 \\ x &= \frac{4.03}{0.66} \approx 6.11 \end{aligned}$$

Question 3: Logistic Regression (20 points)

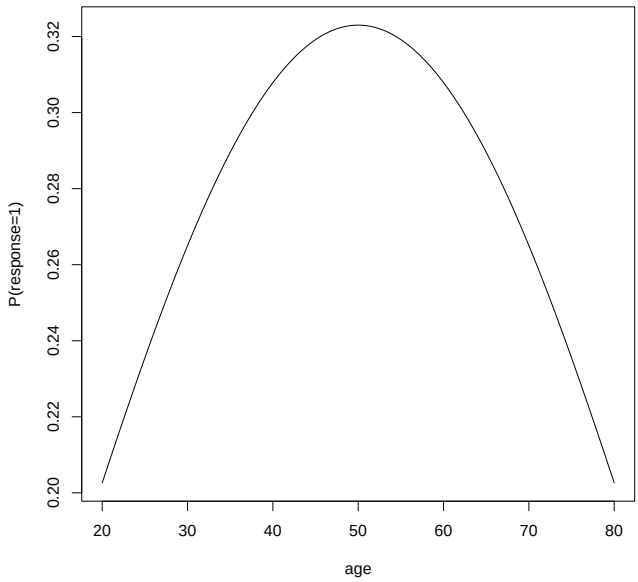
(a) All else equal, men have a higher probability to respond than women. All else equal, people who are already active investors have a higher probability to respond.

(b)

$$z = -2.49 + 0.95 + 30 \times 0.07 - 9 \times 0.07 = -0.07$$

$$\frac{1}{1 + e^{0.07}} \approx 0.4825$$

(c) The probability of response first increases with age, then reaches a peak (at age 50) and then starts to decrease with age. See the graph below, which shows the relationship between age and probability of response for a female who is not yet an active investor.



(d) All of them. All p-values are smaller than 0.05.