Introduction of Machine Learning

Ex2: Classifiers and Bayes Decision Theory

TA: Oscar Wu, oscarwu217@alum.ccu.edu.tw
Lecturer: Prof. Peggy Lu, peggylu@cs.ccu.edu.tw

Problem 1

Consider the Table given below where $\mathbf{x}$ is a two-dimensional vector of independent variables. The first 6 data points are used for learning a model and the next 3 data points are used for validating the model prediction. The parameter y indicates the class $\{0, 1\}$ after least squares analysis.

(a) Find the least square solution if the model is $y = w_1x_1 + w_2x_2 + w_3$.

(b) Find the least square solution if the model is $y = w_1x_1 + w_2x_2$.

Point	Type	x_1	x_2	Y-Class
1	Learning	0.5	0	0
2	Learning	0.75	0.25	0
3	Learning	1	0.5	0
4	Learning	0	0.5	1
5	Learning	0.25	0.75	1
6	Learning	0.5	1	1
7	Testing	0.5	0.25	-
8	Testing	0.5	0.75	-
9	Testing	0.75	0.5	-

Table 1: Dataset for Problem 1

Problem 2

To play tennis, it is important to consider the weather conditions. Using a *naïve Bayes classifier*, classify the days based on whether somebody plays on that day or not using the following Table.

Use the training data from the Table to compute the probability of the following new instance:

$$\text{Outlook} = \text{sunny}, \quad \text{Temperature} = \text{cool}, \quad \text{Humidity} = \text{high}, \quad \text{Wind} = \text{strong}.$$

Problem 3

As a result of a dichotomous classification of diseases, the patient might have heart disease (1) or not (0). We want to study the effect of smoking on the heart disease problem. The additional independent variables which enter the problem are race, sex, and three other health conditions $x_1 = CAT$, $x_2 = EGG$, and $x_3 = AGE$. If the age of a person is equal to a, then $x_3 = AGE$ is computed as:

Day	Outlook	Temperature	Humidity	Wind	Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Overcast	Cool	Normal	Weak	Yes
D7	Rain	Cool	Normal	Strong	No
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Table 2: Weather Dataset for Naïve Bayes Classification

$$x_3 = \frac{1}{3} \log a.$$

The logistic function that relates the variables to the disease is:

$$p(x) = \frac{1}{1 + e^{-z}}$$

where $z = a_0 + a_1x_1 + a_2x_2 + a_3x_3$. Information is gathered for 700 white males over 10 years. We suppose that after learning the model, the following results are derived:

(a) $a_0 = 4$, $a_1 = 0.7$, $a_2 = 0.03$, $a_3 = 0.4$. What is the probability with which a 40-year-old person with $CAT = 1$ and $EGG = 0$ is at heart disease risk?

(b) If logarithmic probability $\log p(x) = \ln \frac{p(x)}{1-p(x)}$ represents the odds for a person developing the disease with independent variable x, compute the odds for part 1 of the problem.

(c) If all of x variables are zero, what does $\ln p(x)$ show?