ISMET OKAN CELIK CWID:10472265

Question 1 (6 points):

Prove Bayes' Theorem. Briefly explain why it is useful for machine learning problems.

From given data directly getting a model is not easy, we don't really know which model is most likely. Basically, given the data, we want to find out the hypothesis, and in machine learning, we aim to determine the best hypothesis from hypothesis space. Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability. The probabilities of observing various data given the hypothesis, and observed data itself. That's why Bayes' Theorem is useful for machine learning problems.

$$P(h|D) = \frac{P(D|h)\,P(h)}{P(D)}$$

$P(h)$: Prior probability of hypothesis h
$P(D)$: Prior probability of training data D
$P(h|D)$: Probability of h given D (Posterior probability)
$P(D|h)$: Probability of D given h
D: Training Data
h: Model

Given the data, we want the most probable hypothesis
Maximum a posterior hypothesis h_{MAP} (Given the data this model is best)

$$h_{MAP} = \underset{h \in H}{\arg\max}\ P(h|D)$$

H: Hypothesis Space

$$= \arg\max \frac{P(D|h)\,P(h)}{P(D)}$$

$$= \arg\max\ P(D|h)\,P(h)$$

→ we just need to maximize the numerator.

we assume all the data evenly distributed.
we assume every hypothesis in H is equally probable a prior ($P(h_i) = P(h_j)$ for all h_i an h_j in H)
If we further simplfy to equation and choose the Maximum Likelihood (ML) hypothesis

$$h_{ML} = \underset{h_i \in H}{\arg\max}\ P(D|h_i)$$

Question 2 (8 points):

Consider again the example application of Bayes rule in Section 6.2.1 of Tom Mitchell's textbook. Suppose the doctor decides to order a second laboratory test for the same patient and suppose the second test returns a positive result as well. What are the posterior probabilities of cancer and ¬cancer respectively following these two tests? Assume that the two tests are independent.

$$P(\text{cancer} \mid +) = \frac{P(+ \mid \text{cancer}) \cdot P(\text{Cancer})}{P(+ \mid \text{cancer}) \cdot P(\text{Cancer}) + P(+ \mid \text{not cancer}) \cdot P(\text{not cancer})}$$

$$= \frac{0{,}98 \cdot 0{,}008}{0{,}98 \cdot 0{,}008 + 0{,}992 \cdot 0{,}03} = 0{,}21$$

$$P(\text{Not Cancer} \mid +) = 1 - 0{,}21 = 0{,}79$$

$$P(\text{Cancer} \mid ++) = \frac{0{,}98 \cdot 0{,}21}{0{,}98 \cdot 0{,}21 + 0{,}03 \cdot 0{,}79} = \underline{0{,}90}$$

$$P(\text{not cancer} \mid ++) = 1 - 0{,}90 = \underline{0{,}1}$$

Question 3 (8 points):

Section 6.9.1 of Tom Mitchell's textbook demonstrates an example using the Naïve Bayes Algorithm to predict a new instance based on a dataset with 14 examples from Table 3.2 of Chapter 3 of the book. If we only have 12 examples as shown below, what is the prediction results for the same new instance? Show your calculation.

New instance: <Outlook=sun, Temperature=cool, Humidity=high, Wind=strong>

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
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	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
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

$$P(Yes)\,P(Sun|Yes)\,P(Cool|Yes)\,P(High|Yes)\,P(Strong|Yes) = \frac{8}{12}\cdot\frac{2}{8}\cdot\frac{3}{8}\cdot\frac{3}{8}\cdot\frac{3}{8} = 8{,}79 \times 10^{-3}$$
$$= 0{,}00879$$

$$P(No)\,P(Sun|No)\,P(Cool|No)\,P(High|No)\,P(Strong|No) = \frac{4}{12}\cdot\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{2}{4} = 0{,}0234$$

Question 4 (14 points): Answer question 4.7 (page 125) of Tom Mitchell's textbook as quoted below:

Consider a two-layer feedforward ANN with two inputs a and b, one hidden unit c, and one output unit d. This network has five weights ($w_{ca}, w_{cb}, w_{c0}, w_{dc}, w_{d0}$), where w_{x0} represents the threshold weight for unit x. Initialize these weights to the values (.1, .1, .1, .1, .1), then give their values after each of the first two training iterations of the BACKPROPAGATION algorithm. Assume learning rate $\eta = .3$, momentum $\alpha = 0.9$, incremental weight updates, and the following training examples:

a	b	d
1	0	1
0	1	0

$w_{ca} = 0{,}1$, $w_{cb} = 0{,}1$, $w_{c0} = 0{,}1$, $w_{dc} = 0{,}1$, $w_{d0} = 0{,}1$ $\quad \eta = 0{,}3 \quad \alpha = 0{,}9$

Training Iteration-1

$$C_{output} = \sigma(\underbrace{w_{ca}\cdot a + w_{cb}\cdot b + w_{c0}}_{net}) \rightarrow \sigma() \quad \frac{1}{1+e^{-net}}$$

Step 1)

$$= \sigma(0{,}1\cdot 1 + 0{,}1\cdot 0 + 0{,}1) = \sigma(0{,}2) = 0{,}5498$$

$$d_{output} = \sigma(0{,}1\cdot 0{,}5498 + 0{,}1) = \sigma(0{,}15498) = 0{,}5387$$

Step 2)

$$E_d = d_{output}\cdot(1 - d_{output})\cdot(t_d - d_{output})$$
$$= 0{,}5387(1 - 0{,}5387)(1 - 0{,}5387)$$
$$E_d = 0{,}1146$$

$$E_c = C_{output}\cdot(1 - C_{output})\cdot(w_c\cdot E_d)$$
$$E_c = 0{,}5498(1 - 0{,}5498)\cdot 0{,}1\cdot 0{,}1146$$
$$E_c = 0{,}00284$$

$$\Delta w_{i,j}(n) = \eta\cdot E\cdot x_{i,j} + \alpha\,\Delta w_{i,j}(n-1)$$

Step 3)

$$\Delta w_{ca} = 0{,}3\cdot 0{,}00284\cdot 1 = 0{,}000852$$
$$\Delta w_{cb} = 0{,}3\cdot 0{,}00284\cdot 0 = 0$$
$$\Delta w_{c0} = 0{,}3\cdot 0{,}00284\cdot 1 = 0{,}000852$$

$$\Delta w_{dc} = 0{,}3 \cdot 0{,}1146 \cdot 0{,}5498 = 0{,}0189$$
$$\Delta w_{do} = 0{,}3 \cdot 0{,}1146 \cdot 1 = 0{,}03438$$

Step 4)

$$w_{ca} = 0{,}1 + 0{,}000852 = 0{,}100852$$
$$w_{cb} = 0{,}1 + 0 = 0{,}1$$
$$w_{co} = 0{,}1 + 0{,}000852 = 0{,}100852$$
$$w_{dc} = 0{,}1 + 0{,}0189 = 0{,}1189$$
$$w_{do} = 0{,}1 + 0{,}03438 = 0{,}13438$$

Iteration - 2

Step 1)

$$c_{output-2} = \sigma(0{,}100852 \cdot 1 + 0{,}1 \cdot 0 + 0{,}100852) = \sigma(0{,}2017) = 0{,}5502$$
$$d_{output-2} = \sigma(0{,}1189 \cdot 0{,}55 + 0{,}13438 \cdot 1) = \sigma(0{,}1997) = 0{,}5497$$

Step2)

$$E_d = d_{output-2}(1 - d_{output-2}) \cdot (t_d - d_{output-2})$$
$$= 0{,}5497(1 - 0{,}5497) \cdot (0 - 0{,}5497)$$
$$= -0{,}136$$
$$E_c = C_{output-2}(1 - C_{output-2}) \cdot (w_{dc} \cdot E_d)$$
$$= 0{,}5502(1 - 0{,}5502) \cdot 0{,}1189 \cdot (-0{,}136)$$
$$= -0{,}004$$

Step 3)

$$\Delta w_{dc} = 0{,}3 \cdot (-0{,}136) \cdot 0{,}55 + 0{,}9 \cdot 0{,}0189 = -0{,}00543$$
$$\Delta w_{do} = 0{,}3 \cdot (-0{,}136) \cdot 1 + 0{,}9 \cdot 0{,}03438 = -0{,}01$$
$$\Delta w_{ca} = 0{,}3 \cdot (-0{,}004) \cdot 0 + 0{,}9 \cdot 0{,}000852 = 0{,}00076$$
$$\Delta w_{cb} = 0{,}3 \cdot (-0{,}004) \cdot 1 + 0{,}9 \cdot 0 = -0{,}0012$$
$$w_{wco} = 0{,}3 \cdot (-0{,}004) \cdot 1 + 0{,}9 \cdot 0{,}000852 = -0{,}0004$$

Step 4)

$$w_{dc} = 0{,}1189 + (-0{,}00543) = 0{,}11347$$
$$w_{do} = 0{,}13438 + (-0{,}01) = 0{,}12438$$
$$w_{ca} = 0{,}100852 + 0{,}00076 = 0{,}10161$$
$$w_{cb} = 0{,}1 - 0{,}0012 = 0{,}0988$$
$$w_{co} = 0{,}100852 - 0{,}0004 = 0{,}100452$$