Q1 Naive Bayes Classifier

34 Points

Consider the tennis dataset below.

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Predict the class Play for the new example: (Outlook = rainy, Temperature = cool, Humidity = high, Windy =false)

No need to calculate all probabilities, calculate only the ones you need to make this prediction. Use simple probabilities (no m-estimate or smoothing). Show all your work.

Please show your work.

Only do one of the answer options, a text entry or a screenshot of latex, not both. (If you do both we will grade only the text entry.)

▾ 1.png Download

We have a total of fourteen data points.

$$\Pr(\text{yes}) = \frac{9}{14},$$
$$\Pr(\text{no}) = \frac{5}{14}.$$

From the table, we know that the conditional probabilities of **Outlook, Temperature, Humidity and Windy** providing **yes** or **no** are

$$\Pr(\text{rainy}|\text{yes}) = \frac{3}{9},\ \Pr(\text{rainy}|\text{no}) = \frac{2}{5}$$
$$\Pr(\text{cool}|\text{yes}) = \frac{3}{9},\ \Pr(\text{cool}|\text{no}) = \frac{1}{5}$$
$$\Pr(\text{high}|\text{yes}) = \frac{3}{9},\ \Pr(\text{high}|\text{no}) = \frac{4}{5}$$
$$\Pr(\text{false}|\text{yes}) = \frac{6}{9},\ \Pr(\text{false}|\text{no}) = \frac{2}{5}$$

According to equation

$$y_{\text{new}} = \text{argmax}_{y \in \mathbb{Y}} \Pr(y) \prod_j \Pr(a_j \mid y), \quad (7)$$

where y can be **yes** or **no** for this question, we know that

$$\Pr(\text{yes}) * \Pr(\text{rainy}|\text{yes}) * \Pr(\text{cool}|\text{yes}) * \Pr(\text{high}|\text{yes}) * \Pr(\text{false}|\text{yes}) = \frac{9}{14}\frac{3}{9}\frac{3}{9}\frac{3}{9}\frac{6}{9} \approx 0.016, \quad (8)$$

and

$$\Pr(\text{no}) * \Pr(\text{rainy}|\text{no}) * \Pr(\text{cool}|\text{no}) * \Pr(\text{high}|\text{no}) * \Pr(\text{false}|\text{no}) = \frac{5}{14}\frac{2}{5}\frac{1}{5}\frac{4}{5}\frac{2}{5} \approx 0.009. \quad (9)$$

Therefore, for the **new** example with **(rainy, cool, high, false)**, we predict y_{new} =**yes**.

Q2 Perceptron

30 Points

Consider the dataset below on which we ran the perceptron algorithm. We report in this table the number of times each example was mis-classified during the run.

x_1	x_2	x_3	y	#times misclassified
2	3	1	+1	12
2	4	0	+1	0
3	1	1	-1	3
1	1	0	-1	6
1	2	1	-1	11

Assume the learning rate is 1, and the initial weights are all zeros.

Q2.1
18 Points

Find the weights, (bias, w_1, w_2, w_3), of the separating line then write the equation of the separating line in function of x_1, x_2, x_3? Don't forget to add a feature x_0 set to 1 for all examples, as seen in class. Show all your calculations.

Please show your work.

Only do one of the answer options, a text entry or a screenshot of latex, not both. (If you do both we will grade only the text entry.)

2.1.png Download

The initial classifier was

$$f(x) = \text{sign}(\omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3), \tag{10}$$

where all weights were set to ZERO, with x_0 being ONE for all examples. We denote any example to be $(x_0 = 1, x_1, x_2, x_3, y)$, i.e., $(1, x_1, x_2, x_3, y)$. Since we misclassified $(1, 2, 3, 1, +1)$ 12 times, $(1, 3, 1, 1, -1)$ 3 times, $(1, 1, 1, 0, -1)$ 6 times and $(1, 1, 2, 1, -1)$ 11 times. According to the rule of the Perceptron algorithm

$$w_j := w_j + y_i x_{ij}, \tag{11}$$

where $j = 0, 1, 2, 3$, our weights were updated as

$$\begin{aligned}
\omega_0 &= 0 + 1 \times 12 - 1 \times 3 - 1 \times 6 - 1 \times 11 = -8, \\
\omega_1 &= 0 + 2 \times 12 - 3 \times 3 - 1 \times 6 - 1 \times 11 = -2, \\
\omega_2 &= 0 + 3 \times 12 - 1 \times 3 - 1 \times 6 - 2 \times 11 = +5, \\
\omega_3 &= 0 + 1 \times 12 - 1 \times 3 - 0 \times 6 - 1 \times 11 = -2.
\end{aligned}$$

Therefore, our classifier became

$$f(x) = \text{sign}(-8 - 2x_1 + 5x_2 - 2x_3), \tag{12}$$

where the weights, $(\text{bias}, \omega_1, \omega_2, \omega_3)$ is $(-8, -2, +5, -2)$.

Q2.2
12 Points

How would the perceptron result change if we add the additional training point (1,1,1) with label +1? Justify your answer (you can plot the data or use other justification).

Please show your work.

Only do one of the answer options, a text entry or a screenshot of latex, not both. (If you do both we will grade only the text entry.)

2.2-1.png Download

Put the values of the features of the new example into the current classifier

$$f(x) = \text{sign}(-8 - 2x_1 + 5x_2 - 2x_3), \tag{13}$$

we have

$$f(x) = \text{sign}(-8 - 2 \times 1 + 5 \times 1 - 2 \times 1) = \text{sign}(-7) = -1 \neq +1. \tag{14}$$

That is, we made a mistake. So we want to update the classifier as we did in the first part of this question

$$\begin{aligned}
\omega_0 &= -8 + 1 \times 1 = -7, \\
\omega_1 &= -2 + 1 \times 1 = -1, \\
\omega_2 &= +5 + 1 \times 1 = +6, \\
\omega_3 &= -2 + 1 \times 1 = -1.
\end{aligned}$$

Our new classifier is now

$$f(x) = \text{sign}(-7 - x_1 + 6x_2 - x_3). \tag{15}$$

Now, the prediction for this new example is

$$f(x) = \text{sign}(-7 - x_1 + 6x_2 - x_3) = \text{sign}(-7 - 1 + 6 \times 1 - 1) = \text{sign}(-3) = -1, \tag{16}$$

which is still wrong. For other examples in the table, we have **(we will still use the notation we defined in the first part of this question)**

1. (1,2,3,1,+1), we predict +1. And this is correct.
2. (1,2,4,0,+1), we predict +1. And this is correct.
3. (1,3,1,1,-1), we predict -1. And this is correct.
4. (1,1,1,0,-1), we predict -1. And this is correct.
5. (1,1,2,1,-1), we predict +1. And this is **wrong**.

So, our classifier cannot correctly predict examples in the dataset. From the figure, we know that the added new example voided the Perceptron algorithm, as it made the dataset not linearly spreadable. So the algorithm will never output a valid classifier that is consistent with the new dataset.

2.2-2.png Download

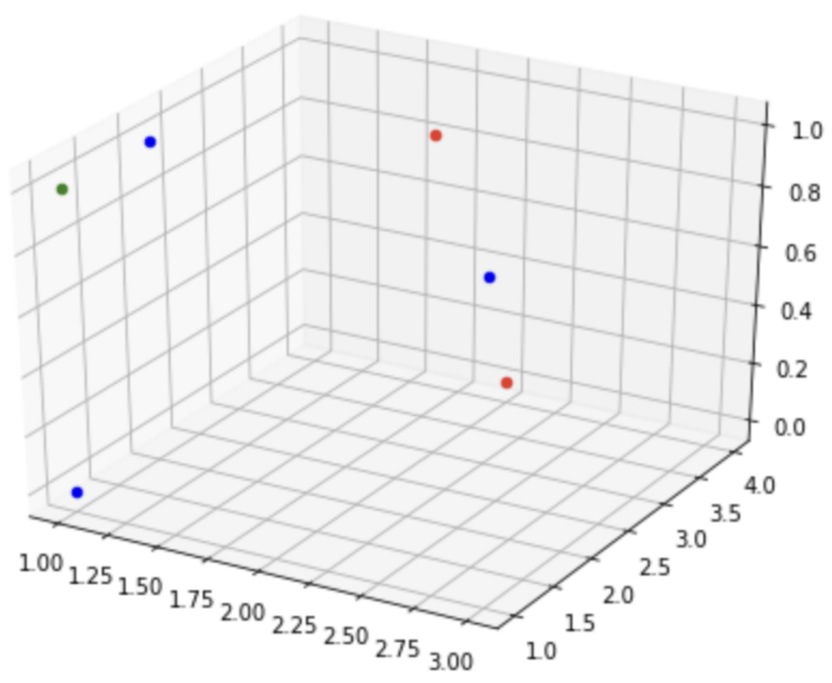

Figure 2: In this figure, red points denote positive examples, whereas blue points denote negative examples. The green point is the new example added to the dataset.

Q3 Ensemble Models

36 Points

We want to build an ensemble model using majority voting of N independent classifiers. Let N an odd number. Suppose each classifier makes an error ε < 0.5.

Q3.1

6 Points

In what cases does the ensemble make a prediction error? Explain.

Please show your work.

Only do one of the answer options, a text entry or a screenshot of latex, not both. (If you do both we will grade only the text entry.)

3.1.png Download

Since we have N independent classifiers and we use the majority voting method to determine the final results, we will get a prediction error if more than half of the classifiers in the pool do not make the correct prediction. Provided N is an odd number, then we can say that if at least $\frac{N+1}{2}$ classifiers output the wrong prediction, then we make a prediction error.

Q3.2

8 Points

What is the error of the ensemble algorithm in function of N and ε? Show all your work.

Please show your work.

Only do one of the answer options, a text entry or a screenshot of latex, not both. (If you do both we will grade only the text entry.)

3.2.png Download

The probability of k classifiers out of N classifiers incorrectly making the prediction is

$$\binom{N}{k}\epsilon^k(1-\epsilon)^{N-k}. \tag{17}$$

If at least $\frac{N+1}{2}$ classifiers incorrectly make predictions, and we cannot get the correct prediction. Then we get the following equation

$$\sum_{k=\frac{N+1}{2}}^{N}\binom{N}{k}\epsilon^k(1-\epsilon)^{N-k}, \tag{18}$$

which is the desired error of the ensemble method in function of N and ϵ.

Q3.3

8 Points

Calculate the error in the case of ε=0.4 for N=5, N=11, and N=21. Compare the error of the ensemble to ε. What do you observe as N increases? Show all your work.

Please show your work.

Only do one of the answer options, a text entry or a screenshot of latex, not both. (If you do both we will grade only the text entry.)

3.3.png Download

According to the equation in the second part of this question, we just put $\epsilon = 0.4$ and $N = 5, 11, 21$ into the function. **Then we have the error rate to be 31.7%, 24.7% and 17.4% for $N = 5$, $N = 11$ and $N = 21$, respectively.**

A Python script was used to make the calculation as follows

```
import math
def nCr(n, r):
    f = math.factorial
    return f(n) // f(r) // f(n-r)

def calculate_error(epsilon, N):
    half = (N + 1) // 2
    sum = 0
    for i in range(half, N + 1):
        curP = nCr(N, i) * math.pow(epsilon, i) *
                           math.pow(1 - epsilon, N - i)
        sum += curP
    return sum

error = calculate_error(0.4, 5)
print(error) # 0.3174400000000001

error = calculate_error(0.4, 11)
print(error) # 0.24650186752000006

error = calculate_error(0.4, 21)
print(error) # 0.17437786636177277
```

That is, the error of the ensemble method is better than individual classifier. And the error is minimised if we have more independent classifiers.

Q3.4

6 Points

In the case that ε > 0.5, would the ensemble perform worse or better than ε? Justify your work with lecture content or a numerical example.

Please show your work.

Only do one of the answer options, a text entry or a screenshot of latex, not both. (If you do both we will grade only the text entry.)

3.4.png Download

Now we set $\epsilon = 0.6$ and try $N = 5, 11, 21$.

```
import math
def nCr(n, r):
    f = math.factorial
    return f(n) // f(r) // f(n-r)

def calculate_error(epsilon, N):
    half = (N + 1) // 2
    sum = 0
    for i in range(half, N + 1):
        curP = nCr(N, i) * math.pow(epsilon, i) *
                           math.pow(1 - epsilon, N - i)
        sum += curP
    return sum

error = calculate_error(0.6, 5)
print(error) # 0.68256

error = calculate_error(0.6, 11)
print(error) # 0.7534981324799999

error = calculate_error(0.6, 21)
print(error) # 0.8256221336382271
```

Using the Python code above, **we can get error rates 68.3%, 75.3% and 82.6% for $N = 5$, $N = 11$ and $N = 21$, respectively. That is, the ensemble method performs worse than an individual classifier. And the more individual classifiers we have, the higher the error of the ensemble method is.**

We may notice that when we set $\epsilon = 0.6$, the error rates for $N = 5, 11, 21$ are equal to the ONE minus the error rates for $\epsilon = 0.4$ and $N = 5, 11, 21$, respectively. This is because we are literally calculating the opposite situations. And intuitively, with more BAD classifiers, we cannot get GOOD results if we simply combine them.

Q3.5

8 Points

In the case that the classifiers used are not independent, would the ensemble perform worse or better than ε? Justify your work with a numerical justification. Suggestion: use N = 5, ε = 0.4, and five examples. Other numerical justifications will be accepted. Show all your work.

Please show your work.

Only do one of the answer options, a text entry or a screenshot of latex, not both. (If you do both we will grade only the text entry.)

No files uploaded

HW4 Conceptual

UNGRADED

STUDENT
Ziggy Chen

TOTAL POINTS
- / 100 pts

QUESTION 1
Naive Bayes Classifier — 34 pts

QUESTION 2
Perceptron — 30 pts

2.1 (no title) — 18 pts

2.2 (no title) — 12 pts

QUESTION 3
Ensemble Models — 36 pts

3.1 (no title) — 6 pts

3.2 (no title) — 8 pts

3.3 (no title) — 8 pts

3.4 (no title) — 6 pts

3.5 (no title) — 8 pts