CSCI 347: Introduction to Data Mining

Introduction to Classification

CLASSIFICATION

Class label is always categorical

Input Data Matrix:

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

New Data Instance:

Weather	Weekend?	Finished HW?	Go Hiking?
Sunny	Yes	No	?

Goal: Predict the class of new data

CLASSIFICATION

Input also commonly in the form:

Weather	Weekend?	Finished HW?	"Target/ Label/Class"
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

New Data Instance:

Weather	Weekend?	Finished HW?	Target/Label/ Class
Sunny	Yes	No	?

CLASSIFICATION

Input also commonly in the form:

Weather	Weekend?	Finished HW?
Snow	Yes	No
Overcast	No	No
Sunny	Yes	No
Overcast	Yes	Yes
Overcast	No	Yes
Snow	No	Yes
Overcast	Yes	No
Sunny	Yes	No
Sunny	No	Yes
Snow	No	Yes
Snow	Yes	No
Overcast	Yes	No
Overcast	No	Yes

y
Yes
No
Yes
Yes
Yes
No
No
No
Yes
Yes
No
No
Yes

New Data Instance:

Weather	Weekend?	Finished HW?	y
Sunny	Yes	No	y=?

CSCI 347: Introduction to Data Mining

Introduction to Naive Bayes Algorithm

NAIVE BAYES IS A CLASSIFICATION ALGORITHM

Input Data Matrix:

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

New Data Instance:

Weather	Weekend?	Finished HW?	Go Hiking?
Sunny	Yes	No	?

NAIVE BAYES IS A CLASSIFICATION ALGORITHM

New Data Instance:

Weather	Weekend?	Finished HW?	Go Hiking?
Sunny	Yes	No	?

Input Data Matrix:

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

NAIVE BAYES IS A CLASSIFICATION ALGORITHM

New Data Instance:

Weather	Weekend?	Finished HW?	Go Hiking?
Sunny	Yes	No	?

Input Data Matrix:

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$n_1 = 7$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$n_2 = 6$

NAIVE BAYES IS A CLASSIFICATION ALGORITHM

New Data Instance:

Weather	Weekend?	Finished HW?	Go Hiking?
Sunny	Yes	No	?

What is the probability of "Yes" and what is the probability of "No" in our data? (these are the *prior* probabilities)

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

We want to now estimate the probabilities of c_1 and c_2 ***after observing the new data instance*** *and then* ***choose the one with maximum probability***

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$n_1 = 7$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$n_2 = 6$

NAIVE BAYES FINDS THE CLASS WITH HIGHEST PROBABILITY

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

What is the probability of "Yes" and what is the probability of "No" in **after observing the new data instance?**

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_1 \mid x) = ?$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(c_2 \mid x) = ?$$

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$n_1 = 7$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$n_2 = 6$

NAIVE BAYES USES BAYES' RULE

What is the probability of "Yes" and what is the probability of "No" in after observing the new data instance? **Use Bayes' Rule:**

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_1 \mid x) = \frac{p(x \mid c_1)p(c_1)}{p(x)}$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(c_2 \mid x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$n_1 = 7$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$n_2 = 6$

NAIVE BAYES USES BAYES' RULE TO FIND CLASS PROBABILITIES

Since we are going to choose the c_i that **maximizes** $p(c_i | x)$, we can ignore $p(x)$ and only need to further compute $p(x | c_i)$ **for each class**

$$argmax_{c_i} p(c_i | x) = argmax_{c_i} \frac{p(x | c_i) p(c_i)}{p(x)}$$

$$= argmax_{c_i} p(x | c_i) p(c_i)$$

$$p(c_1 | x) = \frac{p(x | c_1) p(c_1)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_2 | x) = \frac{p(x | c_2) p(c_2)}{p(x)}$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$n_1 = 7$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$n_2 = 6$

NAIVE BAYES USES BAYES' RULE TO FIND CLASS PROBABILITIES

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

Since we are going to choose the c_i that **maximizes** $p(c_i \mid x)$, we can ignore $p(x)$ and only need to further compute $p(x \mid c_i)$ **for each class**

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$n_1 = 7$

$$argmax_{c_i} p(c_i \mid x) = argmax_{c_i} \frac{p(x \mid c_i)p(c_i)}{p(x)}$$

$$= argmax_{c_i} p(x \mid c_i)p(c_i)$$

$$p(c_1 \mid x) = \frac{p(x \mid c_1)p(c_1)}{p(x)} \qquad p(c_2 \mid x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(x \mid c_1) = p(X_1 = Sunny, X_2 = Yes, X_3 = No \mid c_1)$$

$$= p(X_1 = Sunny \mid c_1)p(X_2 = Yes \mid c_1)p(X_3 = No \mid c_1)$$

NAIVE BAYES ASSUMES ATTRIBUTES ARE INDEPENDENT

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

We make the **naive** assumption that

$$p(X_1 = \textbf{sunny}, X_2 = \textbf{Yes}, X_3 = \textbf{No} \,|\, c_1)$$

is equivalent to :

$$p(X_1 = \textbf{Sunny} \,|\, c_1) p(X_2 = \textbf{Yes} \,|\, c_1) p(X_3 = \textbf{No} \,|\, c_1)$$

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$n_1 = 7$

$$argmax_{c_i} p(c_i \,|\, x) = argmax_{c_i} \frac{p(x \,|\, c_i) p(c_i)}{p(x)}$$

$$= argmax_{c_i} p(x \,|\, c_i) p(c_i)$$

$$p(c_1 \,|\, x) = \frac{p(x \,|\, c_1) p(c_1)}{p(x)} \qquad p(c_2 \,|\, x) = \frac{p(x \,|\, c_2) p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

..because these probabilities are easer to estimate

$$p(x \,|\, c_1) = p(X_1 = Sunny, X_2 = Yes, X_3 = No \,|\, c_1)$$

$$= p(X_1 = Sunny \,|\, c_1) p(X_2 = Yes \,|\, c_1) p(X_3 = No \,|\, c_1)$$

NAIVE BAYES USES BAYES' RULE TO FIND CLASS PROBABILITIES

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

$$argmax_{c_i} p(c_i | x) = argmax_{c_i} p(x | c_i) p(c_i)$$

$$p(c_1 | x) = \frac{p(x | c_1) p(c_1)}{p(x)} \qquad p(c_2 | x) = \frac{p(x | c_2) p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$n_1 = 7$

$$p(x | c_1) = p(X_1 = Sunny | c_1) p(X_2 = Yes | c_1) p(X_3 = No | c_1)$$

$$p(X_1 = Sunny | c_1) = \frac{2}{7} = 0.29$$

NAIVE BAYES USES BAYES' RULE TO FIND CLASS PROBABILITIES

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} p(x|c_i)p(c_i)$$

$$p(c_1|x) = \frac{p(x|c_1)p(c_1)}{p(x)} \qquad p(c_2|x) = \frac{p(x|c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$n_1 = 7$

$$p(x|c_1) = p(X_1 = Sunny|c_1)p(X_2 = Yes|c_1)p(X_3 = No|c_1)$$

$$p(X_1 = Sunny|c_1) = \frac{2}{7} = 0.29$$

$$p(X_2 = Yes|c_1) = \frac{3}{7} = 0.43$$

NAIVE BAYES USES BAYES' RULE TO FIND CLASS PROBABILITIES

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} p(x|c_i)p(c_i)$$

$$p(c_1|x) = \frac{p(x|c_1)p(c_1)}{p(x)} \qquad p(c_2|x) = \frac{p(x|c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$n_1 = 7$

$$p(x|c_1) = p(X_1 = Sunny|c_1)p(X_2 = Yes|c_1)p(X_3 = No|c_1)$$

$$p(X_1 = Sunny|c_1) = \frac{2}{7} = 0.29$$

$$p(X_2 = Yes|c_1) = \frac{3}{7} = 0.43$$

$$p(X_3 = No|c_1) = \frac{2}{7} = 0.29$$

NAIVE BAYES USES BAYES' RULE TO FIND CLASS PROBABILITIES

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

$$argmax_{c_i} p(c_i \mid x) = argmax_{c_i} p(x \mid c_i) p(c_i)$$

$$p(c_1 \mid x) = \frac{p(x \mid c_1) p(c_1)}{p(x)} \qquad p(c_2 \mid x) = \frac{p(x \mid c_2) p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$n_1 = 7$

$$p(X_1 = Sunny \mid c_1) = \frac{2}{7} = 0.29$$

$$p(X_2 = Yes \mid c_1) = \frac{3}{7} = 0.43$$

$$p(X_3 = No \mid c_1) = \frac{2}{7} = 0.29$$

$$p(x \mid c_1) = \left(\frac{2}{7}\right)\left(\frac{3}{7}\right)\left(\frac{2}{7}\right) = 0.035$$

NAIVE BAYES USES BAYES' RULE TO FIND CLASS PROBABILITIES

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

$$argmax_{c_i} p(c_i | x) = argmax_{c_i} p(x | c_i)p(c_i)$$

$$p(c_1 | x) = \frac{p(x | c_1)p(c_1)}{p(x)} \qquad p(c_2 | x) = \frac{p(x | c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$n_1 = 7$

$$p(X_1 = Sunny | c_1) = \frac{2}{7} = 0.29$$

$$p(X_2 = Yes | c_1) = \frac{3}{7} = 0.43$$

$$p(X_3 = No | c_1) = \frac{2}{7} = 0.29$$

$$p(x | c_1) = \left(\frac{2}{7}\right)\left(\frac{3}{7}\right)\left(\frac{2}{7}\right) = 0.035$$

$$p(x | c_1)p(c_1) = 0.035(0.54) = 0.0189$$

NAIVE BAYES USES BAYES' RULE TO FIND CLASS PROBABILITIES

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

$$argmax_{c_i} p(c_i | x) = argmax_{c_i} p(x | c_i)p(c_i)$$

$$p(c_1 | x) = \frac{p(x | c_1)p(c_1)}{p(x)}$$

$$p(c_2 | x) = \frac{p(x | c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(X_1 = Sunny | c_1) = \frac{2}{7} = 0.29$$

$$p(X_2 = Yes | c_1) = \frac{3}{7} = 0.43$$

$$p(X_2 = No | c_1) = \frac{2}{7} = 0.29$$

$$p(x | c_1) = \left(\frac{2}{7}\right)\left(\frac{3}{7}\right)\left(\frac{2}{7}\right) = 0.035$$

$$p(x | c_1)p(c_1) = 0.0189$$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$$n_2 = 6$$

$$p(X_1 = Sunny | c_2) = \frac{1}{6} = 0.17$$

NAIVE BAYES USES BAYES' RULE TO FIND CLASS PROBABILITIES

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

$$argmax_{c_i} p(c_i | x) = argmax_{c_i} p(x | c_i) p(c_i)$$

$$p(c_1 | x) = \frac{p(x | c_1) p(c_1)}{p(x)}$$

$$p(c_2 | x) = \frac{p(x | c_2) p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(X_1 = Sunny | c_1) = \frac{2}{7} = 0.29$$

$$p(X_1 = Sunny | c_2) = \frac{1}{6} = 0.17$$

$$p(X_2 = Yes | c_1) = \frac{3}{7} = 0.43$$

$$p(X_2 = No | c_1) = \frac{2}{7} = 0.29$$

$$p(x | c_1) = \left(\frac{2}{7}\right)\left(\frac{3}{7}\right)\left(\frac{2}{7}\right) = 0.035$$

$$p(x | c_1) p(c_1) = 0.0189$$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$$n_2 = 6$$

$$p(X_2 = Yes | c_2) = \frac{4}{6} = 0.67$$

NAIVE BAYES USES BAYES' RULE TO FIND CLASS PROBABILITIES

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

$$argmax_{c_i} p(c_i | x) = argmax_{c_i} p(x | c_i) p(c_i)$$

$$p(c_1 | x) = \frac{p(x | c_1) p(c_1)}{p(x)}$$

$$p(c_2 | x) = \frac{p(x | c_2) p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(X_1 = Sunny | c_1) = \frac{2}{7} = 0.29$$

$$p(X_1 = Sunny | c_2) = \frac{1}{6} = 0.17$$

$$p(X_2 = Yes | c_1) = \frac{3}{7} = 0.43$$

$$p(X_2 = Yes | c_2) = \frac{4}{6} = 0.67$$

$$p(X_2 = No | c_1) = \frac{2}{7} = 0.29$$

$$p(x | c_1) = \left(\frac{2}{7}\right)\left(\frac{3}{7}\right)\left(\frac{2}{7}\right) = 0.035$$

$$p(x | c_1) p(c_1) = 0.0189$$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$$n_2 = 6$$

$$p(X_3 = No | c_2) = \frac{5}{6} = 0.83$$

NAIVE BAYES USES BAYES' RULE TO FIND CLASS PROBABILITIES

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} p(x|c_i)p(c_i)$$

$$p(c_1|x) = \frac{p(x|c_1)p(c_1)}{p(x)}$$

$$p(c_2|x) = \frac{p(x|c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(X_1 = Sunny|c_1) = \frac{2}{7} = 0.29$$

$$p(X_1 = Sunny|c_2) = \frac{1}{6} = 0.17$$

$$p(X_2 = Yes|c_1) = \frac{3}{7} = 0.43$$

$$p(X_2 = Yes|c_2) = \frac{4}{6} = 0.67$$

$$p(X_2 = No|c_1) = \frac{2}{7} = 0.29$$

$$p(X_3 = No|c_2) = \frac{5}{6} = 0.83$$

$$p(x|c_1) = \left(\frac{2}{7}\right)\left(\frac{3}{7}\right)\left(\frac{2}{7}\right) = 0.035$$

$$p(x|c_1)p(c_1) = 0.0189$$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$$n_2 = 6$$

$$p(x|c_2) = \left(\frac{1}{6}\right)\left(\frac{4}{6}\right)\left(\frac{5}{6}\right) = 0.093$$

NAIVE BAYES USES BAYES' RULE TO FIND CLASS PROBABILITIES

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	?

$$argmax_{c_i} p(c_i | x) = argmax_{c_i} p(x | c_i)p(c_i)$$

$$p(c_1 | x) = \frac{p(x | c_1)p(c_1)}{p(x)} \qquad p(c_2 | x) = \frac{p(x | c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$n_2 = 6$

$$p(X_1 = Sunny | c_1) = \frac{2}{7} = 0.29 \qquad p(X_1 = Sunny | c_2) = \frac{1}{6} = 0.17$$

$$p(X_2 = Yes | c_1) = \frac{3}{7} = 0.43 \qquad p(X_2 = Yes | c_2) = \frac{4}{6} = 0.67$$

$$p(X_2 = No | c_1) = \frac{2}{7} = 0.29 \qquad p(X_3 = No | c_2) = \frac{5}{6} = 0.83$$

$$p(x | c_1) = \left(\frac{2}{7}\right)\left(\frac{3}{7}\right)\left(\frac{2}{7}\right) = 0.035 \qquad p(x | c_2) = \left(\frac{1}{6}\right)\left(\frac{4}{6}\right)\left(\frac{5}{6}\right) = 0.093$$

$$p(x | c_1)p(c_1) = 0.0189$$

$$p(x | c_2)p(c_2) = (0.093)(0.46) = 0.0428$$

NAIVE BAYES USES BAYES' RULE TO FIND CLASS PROBABILITIES

$$\operatorname{argmax}_{c_i} p(c_i|x) = \operatorname{argmax}_{c_i} p(x|c_i)p(c_i)$$

$$p(c_1|x) = \frac{p(x|c_1)p(c_1)}{p(x)} \qquad p(c_2|x) = \frac{p(x|c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(X_1 = Sunny|c_1) = \frac{2}{7} = 0.29 \qquad p(X_1 = Sunny|c_2) = \frac{1}{6} = 0.17$$

$$p(X_2 = Yes|c_1) = \frac{3}{7} = 0.43 \qquad p(X_2 = Yes|c_2) = \frac{4}{6} = 0.67$$

$$p(X_2 = No|c_1) = \frac{2}{7} = 0.29 \qquad p(X_3 = No|c_2) = \frac{5}{6} = 0.83$$

$$p(x|c_1) = \left(\frac{2}{7}\right)\left(\frac{3}{7}\right)\left(\frac{2}{7}\right) = 0.035 \qquad p(x|c_2) = \left(\frac{1}{6}\right)\left(\frac{4}{6}\right)\left(\frac{5}{6}\right) = 0.093$$

$$p(x|c_1)p(c_1) = 0.0189 \qquad p(x|c_2)p(c_2) = 0.0428$$

New Data Instance:

	Weather	Weekend?	Finished HW?	Go Hiking?
x	Sunny	Yes	No	No

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$n_1 = 7$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$n_2 = 6$

NAIVE BAYES FOR NUMERICAL ATTRIBUTES

New Data Instance:

x

Inches of Rain in past 2 hours	Hours of sleep previous night	Percentage HW Finished	Go Hiking?
0.2	7.5	85	?

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} p(x|c_i)p(c_i)$$

$$p(c_1|x) = \frac{p(x|c_1)p(c_1)}{p(x)} \qquad p(c_2|x) = \frac{p(x|c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(x|c_1) = ? \qquad p(x|c_2) = ?$$

Inches of Rain in Past 2 hours	Hours of sleep previous night	Percentage HW Finished	Go Hiking?
0	9	80	Yes
0.5	5	90	No
1	7	95	Yes
5	7	100	Yes
0.3	8	100	Yes
0.4	4	100	No
0.1	9	27	No
0	9	50	No
0	8	100	Yes
3	10	98	Yes
6	8	95	No
2.1	8	70	No
1.02	8.5	98	Yes

NAIVE BAYES FOR NUMERICAL ATTRIBUTES

New Data Instance:

x

Inches of Rain in past 2 hours	Hours of sleep previous night	Percentage HW Finished	Go Hiking?
0.2	7.5	85	?

Inches of Rain in Past 2 hours	Hours of sleep previous night	Percentage HW Finished	Go Hiking?
0	9	80	Yes
0.5	5	90	No
1	7	95	Yes
5	7	100	Yes
0.3	8	100	Yes
0.4	4	100	No
0.1	9	27	No
0	9	50	No
0	8	100	Yes
3	10	98	Yes
6	8	95	No
2.1	8	70	No
1.02	8.5	98	Yes

$$argmax_{c_i} p(c_i | x) = argmax_{c_i} p(x | c_i)p(c_i)$$

$$p(c_1 | x) = \frac{p(x | c_1)p(c_1)}{p(x)} \qquad p(c_2 | x) = \frac{p(x | c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(x | c_1) = ? \qquad p(x | c_2) = ?$$

Assume a (multivariate) normal distribution

$$p(x | c_i) = \prod_{j=1}^{d} p(x_j | c_i) = \prod_{j=1}^{d} f(x_j | \hat{\mu}_{ij}, \hat{\sigma}^2_{ij})$$

Where:

$$f(x_j | \hat{\mu}_{ij}, \hat{\sigma}^2_{ij}) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left(\frac{-(x_j - \hat{\mu}_{ij})^2}{2\hat{\sigma}^2_{ij}}\right)$$

NAIVE BAYES FOR NUMERICAL ATTRIBUTES

New Data Instance:

x

Inches of Rain in past 2 hours	Hours of sleep previous night	Percentage HW Finished	Go Hiking?
0.2	7.5	85	?

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} p(x|c_i)p(c_i)$$

$$p(c_1|x) = \frac{p(x|c_1)p(c_1)}{p(x)} \qquad p(c_2|x) = \frac{p(x|c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(x|c_1) = ? \qquad p(x|c_2) = ?$$

$$p(x|c_i) = \prod_{j=1}^{d} p(x_j|c_i) = \prod_{j=1}^{d} f(x_i|\hat{\mu}_{ij}, \hat{\sigma}^2_{ij})$$

$$f(x_j|\hat{\mu}_{ij}, \hat{\sigma}^2_{ij}) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left(\frac{-(x_j - \hat{\mu}_{ij})^2}{2\hat{\sigma}^2_{ij}}\right)$$

Inches of Rain in Past 2 hours	Hours of sleep previous night	Percentage HW Finished	Go Hiking?
0	9	80	Yes
0.5	5	90	No
1	7	95	Yes
5	7	100	Yes
0.3	8	100	Yes
0.4	4	100	No
0.1	9	27	No
0	9	50	No
0	8	100	Yes
3	10	98	Yes
6	8	95	No
2.1	8	70	No
1.02	8.5	98	Yes

NAIVE BAYES ALGORITHM (NUMERICAL DATA)

NaiveBayes($D = \{x_j, y_j\}_{j=1}^{n}$) :

1. for $i = 1,\ldots,k$:
    2. $D_i = \{x_j \mid y_j = c_i, j = 1,\ldots,n\}$
    3. $n_i = |D_i|$
    4. $p(c_i) = \dfrac{n_i}{n}$
    5. $\hat{\mu}_i \leftarrow \dfrac{1}{n_i} \sum_{x_j \in D_i} x_j$
    6. $Z_i \leftarrow D_i - 1.\hat{\mu}_i^T$
    7. For $j = 1,\ldots,d$:
        7. $\hat{\sigma}_{ij}^2 \leftarrow \dfrac{1}{n_i} Z_{ij}^T Z_{ij}$
    8. $\hat{\sigma}_i \leftarrow (\hat{\sigma}_{i1}, \ldots, \hat{\sigma}_{id})^T$
9. Return $p(c_i), \hat{\mu}_i, \hat{\sigma}_i$ for all $i \in \{1,\ldots,k\}$

PredictClass(x and $p(c_i), \hat{\mu}_i, \hat{\sigma}_i$ for all $i = 1,\ldots,k$) :

1. $c \leftarrow \mathrm{argmax}_{c_i}\{p(c_i) \prod_{j=1}^{d} f(x_j \mid \hat{\mu}_{ij}, \sigma_{ij}^2)\}$
2. Return c

CSCI 347: Introduction to Data Mining

Evaluating Classification Algorithms

CLASSIFICATION

Class label is always categorical

Input Data Matrix:

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

New Data Instance:

Weather	Weekend?	Finished HW?	Go Hiking?
Sunny	Yes	No	?

Goal: Predict the class of new data

EVALUATION OF CLASSIFICATION

Input Data Matrix:

Weather	Weekend?	Finished HW?	Go Hiking?
~~Snow~~	~~Yes~~	~~No~~	~~Yes~~
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

Test Data Instance:

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes

Evaluate the class predictions of test data

TRAINING SET AND TEST SET

Test Data Instance:

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes

Input Data Matrix:

Weather	Weekend?	Finished HW?	Go Hiking?
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

Evaluate the class predictions of test data based an algorithm that built a model using only training data

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

TRAINING SET AND TEST SET:

OFTEN 80/20 SPLIT

Input Data Matrix:

	Weather	Weekend?	Finished HW?	Go Hiking?
x_1	Snow	Yes	No	Yes
	Overcast	No	No	No
	Sunny	Yes	No	Yes
x_4	Overcast	Yes	Yes	Yes
	Overcast	No	Yes	Yes
	Snow	No	Yes	No
	Overcast	Yes	No	No
	Sunny	Yes	No	No
	Sunny	No	Yes	Yes
x_{10}	Snow	No	Yes	Yes
	Snow	Yes	No	No
	Overcast	Yes	No	No
	Overcast	No	Yes	Yes

Test Data:

	Weather	Weekend?	Finished HW?	Go Hiking?
x_1	Snow	Yes	No	Yes
x_4	Overcast	Yes	Yes	Yes
x_{10}	Snow	No	Yes	Yes

Evaluate the class predictions of test data based on algorithm that built a model using only training data

Training Data:

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

PIPELINE FOR EVALUATION

Training Data:

Weather	Weekend?	Finished HW?	Go Hiking?
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

Give a classification algorithm training data to use to output a model

("fit" the data)

Test Data:

	Weather	Weekend?	Finished HW?	Go Hiking?
x_1	Snow	Yes	No	Yes
x_4	Overcast	Yes	Yes	Yes
x_{10}	Snow	No	Yes	Yes

Feed the model unlabeled test data to make predictions on

("predict" the test data classes)

	Weather	Weekend?	Finished HW?	Go Hiking?
x_1	Snow	Yes	No	Yes
x_4	Overcast	Yes	Yes	Yes
x_{10}	Snow	No	Yes	No

Compare predicted classes to ground truth classes

EVALUATION METRICS

How to compare predicted classes to ground truth classes?

Predicted Classes:

	Weather	Weekend?	Finished HW?	Go Hiking?
x_1	Snow	Yes	No	Yes
x_4	Overcast	Yes	Yes	Yes
x_{10}	Snow	No	Yes	No

Test Data:

	Weather	Weekend?	Finished HW?	Go Hiking?
x_1	Snow	Yes	No	Yes
x_4	Overcast	Yes	Yes	Yes
x_{10}	Snow	No	Yes	Yes

EVALUATION METRICS

How to compare predicted classes to ground truth classes?

Predicted Classes:

	Weather	Weekend?	Finished HW?	Go Hiking?
$\tilde{x}_1$	Snow	Yes	No	$\hat{y}_1 =$ Yes
$\tilde{x}_2$	Overcast	Yes	Yes	$\hat{y}_2 =$ Yes
$\tilde{x}_3$	Snow	No	Yes	$\hat{y}_3 =$ No

Test Data:

	Weather	Weekend?	Finished HW?	Go Hiking?
$\tilde{x}_1$	Snow	Yes	No	$y_1 =$ Yes
$\tilde{x}_2$	Overcast	Yes	Yes	$y_2 =$ Yes
$\tilde{x}_3$	Snow	No	Yes	$y_3 =$ Yes

Accuracy:

$$\frac{1}{n_T} \sum_{i=1}^{n_T} I(y_i = \hat{y}_i)$$

where:

$I(y_i = \hat{y}_i)$ *is* **1** *if* y_i *and* $\hat{y}_i$ *have the same value, and is* **0** *otherwise*

EVALUATION METRICS

How to compare predicted classes to ground truth classes?

Predicted Classes:

	Weather	Weekend?	Finished HW?	Go Hiking?
$\tilde{x}_1$	Snow	Yes	No	$\hat{y}_1 =$ Yes
$\tilde{x}_2$	Overcast	Yes	Yes	$\hat{y}_2 =$ Yes
$\tilde{x}_3$	Snow	No	Yes	$\hat{y}_3 =$ No

Test Data:

	Weather	Weekend?	Finished HW?	Go Hiking?
$\tilde{x}_1$	Snow	Yes	No	$y_1 =$ Yes
$\tilde{x}_2$	Overcast	Yes	Yes	$y_2 =$ Yes
$\tilde{x}_3$	Snow	No	Yes	$y_3 =$ Yes

Accuracy:

$$\frac{1}{n_T}\sum_{i=1}^{n_T} I(y_i = \hat{y}_i) = \frac{1}{3}(1+1+0) = \frac{2}{3}$$

OTHER EVALUATION METRICS

How to compare predicted classes to ground truth classes?

Predicted Classes:

	Weather	Weekend?	Finished HW?	Go Hiking?
x_1	Snow	Yes	No	Yes
x_4	Overcast	Yes	Yes	Yes
x_{10}	Snow	No	Yes	No

Test Data:

	Weather	Weekend?	Finished HW?	Go Hiking?
x_1	Snow	Yes	No	Yes
x_4	Overcast	Yes	Yes	Yes
x_{10}	Snow	No	Yes	Yes

Contingency-based measures:

- *Precision, recall, F-measure*

For binary classification:

- *TP, TN,FP,FN*
- *Sensitivity, specificity*

Area under ROC Curve