Artificial Intelligence/Machine Learning/Deep Learning:
'Bridging the Skills Gap'

Lesson 6: 'Naïve' Bayes Classifier

We will look at a bunch of classifiers in this course including:

Classification Algorithm	Coding Exercise (Yes/No)	Comment
K-Nearest Neighbors (KNN)	√	
Perceptron – Multi-Layer Perceptron	X	Perceptron is the simplest form of a Neural Network.
Naïve Bayes	√	
Logistics Regression	√	
Support Vector Machines and Kernel Functions	√	
Decision Trees	√	
Neural Networks	√	

Naive Bayes models are fast and simple classification algorithms that are often suitable for very high-dimensional datasets. One application of naïve Bayes is text classification, aiming to assign documents (emails, tweets, posts, news) to one or many categories. One example is email spam/not-spam classification. The idea of naïve Bayes is that spam and not-span emails have a different probability distributions.

For naïve Bayes the data doesn't need to be linearly separable.
Naïve Bayes finds the hyperplane that best differentiates one distribution from another…it DOES NOT find hyperplane that separates + from –
A perceptron separates the data and will not converge when data is not linearly separable

Naive Bayes Classifier

↳ typically used for spam filtering < 'spam' / 'ham'

↳ but also used for other classification applications

- recognize handwritten digits on envelopes
- automatic essay grading
- medical diagnosis
- fraud detection

Bayes Rule:

$$\boxed{P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}}$$

Prior = Probability of A before B occured

probability of A when B has occured → Posterior

→ links probability before an event B has occured to probability before B has occured!

Given: A test point X_{test} (email) with features $\underbrace{x_1, x_2, \dots, x_d}^{\text{words of email}}$

goal → maximize $P(Y = y \mid \underbrace{x_1 = word_1, x_2 = word_2, \dots, x_d = word_d}_{\text{exact email never seen before}})$

'spam' 'ham'

$\xrightarrow{\text{Bayes Rule}}$ maximize $$\frac{P(x_1 = word_1, \dots, x_d = word_d \mid Y = y) \cdot P(Y = y)}{P(x_1 = word_1, \dots, x_d = word_d)}$$

→ given the email is spam what would be the email?

Prior belief spam or ham! before we take evidence into account

$\longrightarrow$ MAXIMIZE $\underbrace{P(x_1 = \text{word}_1 \mid Y = y)}$ $\cdots P(x_d = \text{word}_d \mid Y = y)$

$\cdot P(Y = y)$

Naive Bayes

All words in the email received are independent of each other provided that you know the label y (spam)

|| Naive

but computing is cheap and results are not bad!

ex: Given that email is spam, what is the probability that it contains the word 'VIAGRA'?

Does Naive Bayes overestimate or underestimate the probability of spam?

assume x_1 = VIAGRA, x_2 = KENIA } email with any of these 2 words will be likely SPAM!

BUT unlikely for both words to be in same email.

① $P(x_1 = \text{VIAGRA}, x_2 = \text{KENIA} \mid Y = \text{spam})$

NAIVE BAYES $\longrightarrow$ ② $\underbrace{P(x_1 = \text{VIAGRA} \mid Y = \text{spam}) \cdot P(x_2 = \text{KENIA} \mid Y = \text{spam})}$

$\downarrow$

OVERESTIMATES Probability | $P(②) \gg P(①)$

Several models could explain the data "Evidence"

$\rightarrow$ Pick model that maximizes $P(\text{model} \mid \text{data})$

$\sim P(\text{DATA} \mid \text{model}) \cdot P(\text{model})$

Naive Bayes Summary → linear classifier

1 Continuous Example : 2 classes

- Adults : $a = y$
- children : $c = y$

4 adults , 12 children

training data (h, w, y)

h → Hight, w → weight

1 Compute Prior : $P(a) = \frac{4}{16} = 25\%$

$P(c) = 75\%$

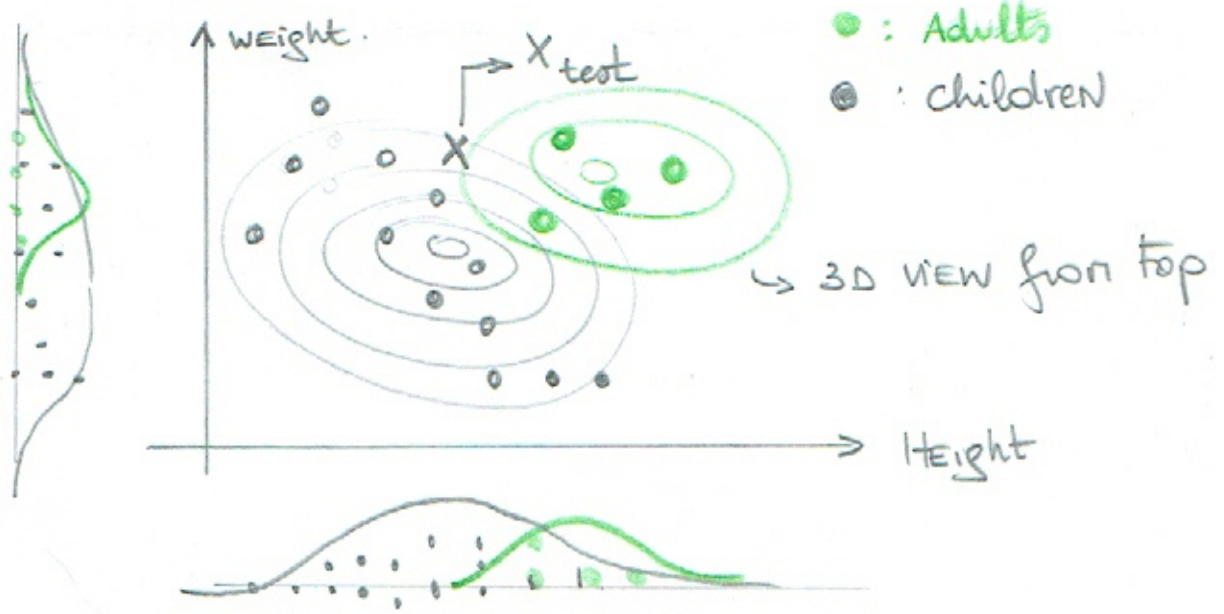
2 model selection : GAUSSIAN (μ, σ^2)

for distribution of weight , hight for adults and for kids } ASSUME iid

$P(x|a)$

$P(x|c)$

DATA Scientist job



$$\text{Distribution} = \underbrace{\text{Distribution}_c}_{\text{GAUSS}} \cdot \underbrace{\text{Distribution}_a}_{\text{GAUSS}}$$

↑ GAUSS!

→ once we have the distributions we can forget about the data.

$$P(x|a) = P(h_x|a)\,P(w_x|a)$$
$$P(x|c) = P(h_x|c)\,P(w_x|c)$$

$$\rightarrow P(a|x) = \frac{P(x|a)\,P(a)}{P(x|a)\,P(a) + P(x|c)\,P(c)}$$

② <u>Discreet Example</u>: Multinomial distribution

<u>GIVEN</u>: Movie Reviews < P (Positive) / N (Negative)

①	bad, terrible, boring	N
②	GOOD, would recommend	P
③	would NOT recommend	N
④	boring, terrible	N

$|V|$ = vocabulary = 7

<u>PRIORS</u>: $P(N) = 3/4 : 75\%$
$P(P) = 1/4 : 25\%$

<u>Conditional Probabilities</u> – likelihoods

$$P(w|c) = \frac{\text{count}(w,c) \boxed{+1}}{\text{count}(c) \boxed{+|V|}}$$

SMOOTHING

↓ TOTAL words in class.

$$P(BAD|N) = \frac{1+1}{8+7} = \frac{2}{15}$$
$$P(\text{terrible}|N) = \frac{2+1}{8+7} = \frac{3}{15}$$
$$P(BORING|N) = \frac{2+1}{8+7} = \frac{3}{15}$$
$$P(Good|N) = \frac{0+1}{8+7} = \frac{1}{15}$$
$$P(\text{would}|N) = \frac{2}{8+7} = \frac{2}{15}$$
$$P(\text{recommend}|N) = \frac{2}{8+7} = 2/15$$
$$P(NOT|N) = \frac{2}{8+7} = 2/15$$

$$P(BAD|P) = \frac{0+1}{3+7} = \frac{1}{10}$$
$$P(\text{Terrible}|P) = \frac{0+1}{3+7} = \frac{1}{10}$$
$$P(BORING|P) = \frac{0+1}{3+7} = \frac{1}{10}$$
$$P(Good|P) = \frac{1+1}{3+7} = 2/10$$
$$P(\text{would}|P) = \frac{1+1}{3+7} = 2/10$$
$$P(\text{recommend}|P) = \frac{2}{3+7} = 2/10$$
$$P(Not|P) = \frac{1}{10}$$

Now we look at a Test Point (Test Review)

5 NOT GOOD, boring "TR

choosing a class:

$$P(N|TR) \sim \frac{3}{4} \cdot \frac{1}{15} \cdot \frac{2}{15} \cdot \frac{3}{15} = \frac{18}{13500} = 0.0013$$

$$= 0.13\%$$

Prior. P(word|N)

↳ Naïve Bayes!

$$P(P|TR) \sim \frac{1}{4} \cdot \frac{1}{10} \cdot \frac{2}{10} \cdot \frac{2}{10} = \frac{4}{4000} = \frac{1}{2000} = 0.1\%$$

↓ Machine will classify Test review as NEGATive because

$0.13 > 0.1$