INTRODUCTION TO MACHINE LEARNING

David Kauchak
CS 51A – Fall 2025

1

Admin

Assignment 6

Assignment 7

Mentor hour change:
- Carmen: Wednesday 8-9pm (used to be 7-8pm)

2

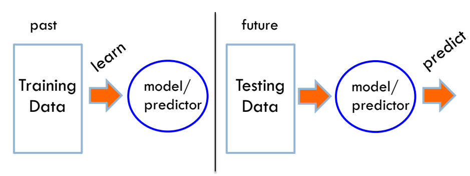
Machine Learning is…

Machine learning is about predicting the future based on the past.
-- Hal Daume III

3

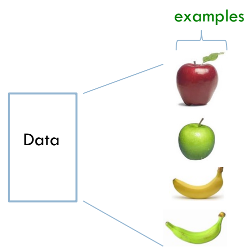
Machine Learning is…

Machine learning is about predicting the future based on the past.
-- Hal Daume III



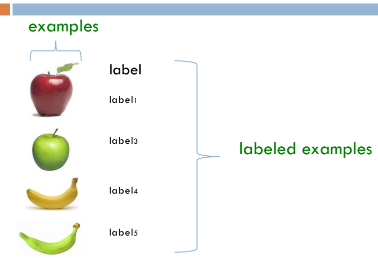
4

Data



5

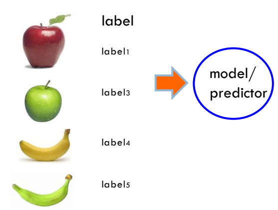
Supervised learning



Supervised learning: given labeled examples

6

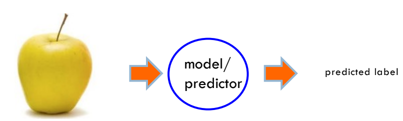
Supervised learning



Supervised learning: given labeled examples

7

Supervised learning



Supervised learning: learn to predict new example

8

Supervised learning: classification

label

apple

apple

banana

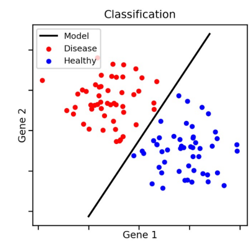
banana

Classification: a finite set of labels

Supervised learning: given labeled examples

9

Classification Example



10

Classification Applications

Face recognition

Character recognition

Spam detection

Medical diagnosis: From symptoms to illnesses

Biometrics: Recognition/authentication using physical and/or behavioral characteristics: Face, iris, signature, etc

...

11

Supervised learning: regression

label

-4.5

10.1

3.2

4.3

Regression: label is real-valued

Supervised learning: given labeled examples

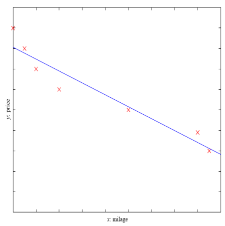
12

Regression Example

Price of a used car

x : car attributes (e.g. mileage)

y : price



13

Regression Applications

Economics/Finance: predict the value of a stock

Epidemiology

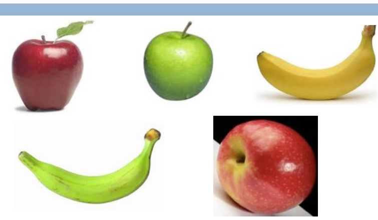
Car/plane navigation: angle of the steering wheel, acceleration, …

Temporal trends: weather over time

…

14

Unsupervised learning



Unsupervised learning: given data, i.e. examples, but no labels

18

Unsupervised learning applications

learn clusters/groups without any label

customer segmentation (i.e. grouping)

image compression

bioinformatics: learn motifs

…

19

Reinforcement learning

left, right, straight, left, left, left, straight	GOOD
left, straight, straight, left, right, straight, straight	BAD

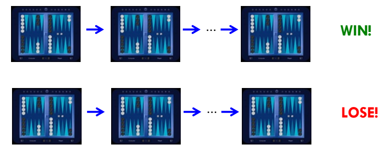
left, right, straight, left, left, left, straight	18.5
left, straight, straight, left, right, straight, straight	-3

Given a *sequence* of examples/states and a *reward* after completing that sequence, learn to predict the action to take in for an individual example/state

20

Reinforcement learning example

Backgammon



Given sequences of moves and whether or not the player won at the end, learn to make good moves

21

Other learning variations

What data is available:

- Supervised, unsupervised, reinforcement learning
- semi-supervised, active learning, …

How are we getting the data:

- online vs. offline learning

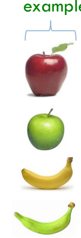
Type of model:

- generative vs. discriminative
- parametric vs. non-parametric

22

Representing examples

examples



What is an example?
How is it represented?

23

Features

examples

features

f1, f2, f3, …, fn

f1, f2, f3, …, fn

f1, f2, f3, …, fn

f1, f2, f3, …, fn

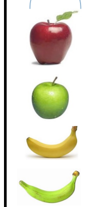
How our algorithms actually "view" the data

Features are the questions we can ask about the examples

24

Features

examples



features

red, round, leaf, 3oz, …

green, round, no leaf, 4oz, …

yellow, curved, no leaf, 8oz, …

green, curved, no leaf, 7oz, …

How our algorithms actually "view" the data

Features are the questions we can ask about the examples

25

Classification revisited

examples	label
red, round, leaf, 3oz, …	apple
green, round, no leaf, 4oz, …	apple
yellow, curved, no leaf, 8oz, …	banana
green, curved, no leaf, 7oz, …	banana

learn → model/classifier

During learning/training/induction, learn a model of what distinguishes apples and bananas *based on the features*

26

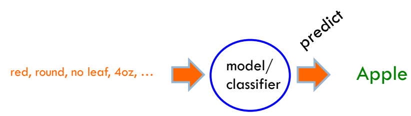
Classification revisited

red, round, no leaf, 4oz, … → model/classifier → predict → Apple or banana?

The model can then classify a new example *based on the features*

27

Classification revisited



Why?

The model can then classify a new example *based on the features*

28

Classification revisited

Training data: examples	label	Test set
red, round, leaf, 3oz, …	apple	
green, round, no leaf, 4oz, …	apple	red, round, no leaf, 4oz, … ?
yellow, curved, no leaf, 4oz, …	banana	
green, curved, no leaf, 5oz, …	banana	

29

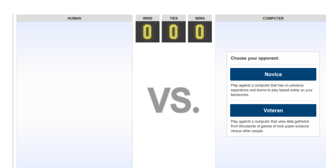
Classification revisited

Training data: examples	label	Test set
red, round, leaf, 3oz, …	apple	
green, round, no leaf, 4oz, …	apple	red, round, no leaf, 4oz, … ?
yellow, curved, no leaf, 4oz, …	banana	Learning is about ***generalizing*** from the training data
green, curved, no leaf, 5oz, …	banana	

30

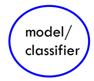
Rock, paper, scissors

https://archive.nytimes.com/www.nytimes.com/interactive/science/rock-paper-scissors.html



31

models

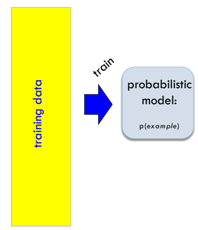


We have many, many different options for the model

They have different characteristics and perform differently (accuracy, speed, etc.)

33

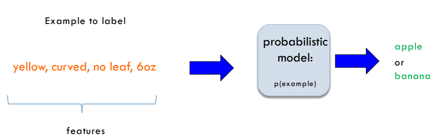
Probabilistic modeling



Model the data with a probabilistic model which tells us how likely a given data example is

34

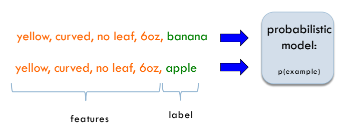
Probabilistic models



35

Probabilistic models

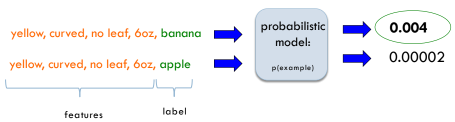
For each label, ask for the probability



36

Probabilistic models

Pick the label with the highest probability



37

Probability basics

A probability distribution gives the probabilities of all possible values of an event

For example, say we flip a coin three times. We can define the probability of the number of time the coin came up heads.

P(num heads)
P(3) = ?
P(2) = ?
P(1) = ?
P(0) = ?

38

Probability distributions

What are the possible outcomes of three flips (hint, there are eight of them)?

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

39

Probability distributions

Assuming the coin is fair, what are our probabilities?

$$\text{probability} = \frac{\text{number of times it happens}}{\text{total number of cases}}$$

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

P(num heads)
P(3) = ?
P(2) = ?
P(1) = ?
P(0) = ?

40

Probability distributions

Assuming the coin is fair, what are our probabilities?

$$\text{probability} = \frac{\text{number of times it happens}}{\text{total number of cases}}$$

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

P(num heads)
P(3) = ?
P(2) = ?
P(1) = ?
P(0) = ?

41

Probability distributions

Assuming the coin is fair, what are our probabilities?

$$\text{probability} = \frac{\text{number of times it happens}}{\text{total number of cases}}$$

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

P(num heads)
P(3) = 1/8
P(2) = ?
P(1) = ?
P(0) = ?

42

Probability distributions

Assuming the coin is fair, what are our probabilities?

$$\text{probability} = \frac{\text{number of times it happens}}{\text{total number of cases}}$$

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

P(num heads)
P(3) = 1/8
P(2) = ?
P(1) = ?
P(0) = ?

43

Probability distributions

Assuming the coin is fair, what are our probabilities?

$$\text{probability} = \frac{\text{number of times it happens}}{\text{total number of cases}}$$

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

P(num heads)
P(3) = 1/8
P(2) = 3/8
P(1) = ?
P(0) = ?

44

Probability distributions

Assuming the coin is fair, what are our probabilities?

$$\text{probability} = \frac{\text{number of times it happens}}{\text{total number of cases}}$$

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

P(num heads)
P(3) = 1/8
P(2) = 3/8
P(1) = 3/8
P(0) = 1/8

45

Probability distribution

A probability distribution assigns probability values to *all possible values*

Probabilities are between 0 and 1, inclusive

The sum of all probabilities in a distribution must be 1

P(num heads)
P(3) = 1/8
P(2) = 3/8
P(1) = 3/8
P(0) = 1/8

46

Probability distribution

A probability distribution assigns probability values to *all possible values*

Probabilities are between 0 and 1, inclusive

The sum of all probabilities in a distribution must be 1

P
P(3) = 1/2
P(2) = 1/2
P(1) = 1/2
P(0) = 1/2

P
P(3) = -1
P(2) = 2
P(1) = 0
P(0) = 0

47

Some example probability distributions

probability of heads
(distribution options: heads, tails)

probability of passing class
(distribution options: pass, fail)

probability of rain today
(distribution options: rain or no rain)

probability of getting an 'A'
(distribution options: A, B, C, D, F)

48

Conditional probability distributions

Sometimes we may know extra information about the world that may change our probability distribution

P(X|Y) captures this (read "probability of X *given* Y")

- Given some information (Y) what does our probability distribution look like
- Note that this is still just a normal probability distribution

49

Conditional probability example

P(pass 51a)
P(pass) = 0.9
P(not pass) = 0.1

Unconditional probability distribution

50

Conditional probability example

P(pass 51a)
P(pass) = 0.9
P(not pass) = 0.1

P(pass 51a | don't study)
P(pass) = 0.5
P(not pass) = 0.5

P(pass 51a | do study)
P(pass) = 0.95
P(not pass) = 0.05

Still probability distributions over passing 51A

Conditional probability distributions

51

Conditional probability example

P(rain in LA)
P(rain) = 0.05
P(no rain) = 0.95

Unconditional probability distribution

52

Conditional probability example

P(rain in LA)
P(rain) = 0.05
P(no rain) = 0.95

P(rain in LA | March)
P(rain) = 0.2
P(no rain) = 0.8

P(rain in LA | not March)
P(pass) = 0.03
P(not pass) = 0.97

Still probability distributions over passing rain in LA

Conditional probability distributions

53

Joint distribution

Probability over two events: P(X,Y)

Has probabilities for all possible combinations over the two events

51Pass, EngPass	P(51Pass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

54

Joint distribution

Still a probability distribution

All questions/probabilities that we might want to ask about these two things can be calculated from the joing distribution

51Pass, EngPass	P(51Pass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

What is P(51pass = true)?

55

Joint distribution

51Pass, EngPass	P(51Pass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

There are two ways that a person can pass 51: they can do it while passing or not passing English

P(51Pass=true) = P(true, true) + P(true, false) = 0.89

56

Relationship between distributions

$$P(X,Y) = P(Y) * P(X|Y)$$

joint distribution — unconditional distribution — conditional distribution

Can think of it as describing the two events happening in two steps:

The likelihood of X and Y happening:
1. How likely it is that Y happened?
2. Given that Y happened, how likely is it that X happened?

57

Relationship between distributions

$$P(51Pass, EngPass) = P(EngPass) * P(51Pass|EngPass)$$

The probability of passing CS51 and English is:
1. Probability of passing English *
2. Probability of passing CS51 **given** that you passed English

58

Relationship between distributions

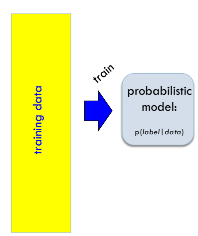
$$P(51Pass, EngPass) = P(51Pass) * P(EngPass|51Pass)$$

The probability of passing CS51 and English is:
1. Probability of passing CS51 *
2. Probability of passing English **given** that you passed CS51

Can also view it with the other event happening first

59

Back to probabilistic modeling



Build a model of the conditional distribution:

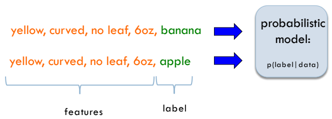
P(label | data)

How likely is a label given the data

60

Back to probabilistic models

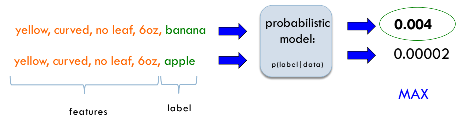
For each label, calculate the probability of the label given the data



61

Back to probabilistic models

Pick the label with the highest probability



62

Naïve Bayes model

Two parallel ways of breaking down the joint distribution

$$P(data, label) = P(label) * P(data|label)$$

$$P(data, label) = P(data) * P(label|data)$$

$$P(label) * P(data|label) = P(data) * P(label|data)$$

What is P(label | data)?

63

Naïve Bayes

$$P(label) * P(data|label) = P(data) * P(label|data)$$

⬇

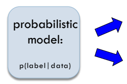
$$P(label|data) = \frac{P(label) * P(data|label)}{P(data)}$$

(This is called Bayes' rule!)

64

Naïve Bayes

$$P(label|data) = \frac{P(label) * P(data|label)}{P(data)}$$



$$\frac{P(positive) * P(data|positive)}{P(data)}$$

$$\frac{P(negative) * P(data|negative)}{P(data)}$$

MAX

65

One observation

$$\frac{P(positive) * P(data|positive)}{P(data)}$$

MAX

$$\frac{P(negative) * P(data|negative)}{P(data)}$$

For picking the largest P(data) doesn't matter!

66

One observation

$$P(positive) * P(data|positive)$$

MAX

$$P(negative) * P(data|negative)$$

For picking the largest P(data) doesn't matter!

67

A simplifying assumption (for this class)

$$P(positive) * P(data|positive)$$

MAX

$$P(negative) * P(data|negative)$$

If we assume P(positive) = P(negative) then:

$$P(data|positive)$$

MAX

$$P(data|negative)$$

68

Naïve Bayes Assumption

$$P(data|label) = P(f_1, f_2, \ldots, fn|label)$$

$$\approx P(f_1\,|label) *$$
$$P(f_2\,|label) *$$
$$\ldots$$
$$P(f_n\,|label)$$

This is generally not true!

However…, it makes our life easier.

This is why the model is called **Naïve** Bayes

69

Naïve Bayes

$P(f_1|positive) * P(f_2|positive) *\ldots* P(f_n|positive)$

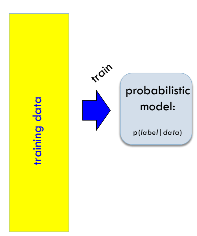
MAX

$P(f_1|negative) * P(f_2|negative) *\ldots* P(f_n|negative)$

Where do these come from?

70

Training Naïve Bayes



71

An aside: P(heads)

What is the P(heads) on a fair coin?

- 0.5

What if you didn't know that, but had a coin to experiment with?

- Flip it a bunch of times and count how many times it comes up heads

$$P(heads) = \frac{number\ of\ times\ heads\ came\ up}{total\ number\ of\ coin\ tosses}$$

72

Try it out…

73

P(feature | label)

$$P(heads) = \frac{number\ of\ times\ heads\ came\ up}{total\ number\ of\ coin\ tosses}$$

Can we do the same thing here? What is the probability of a feature given positive, i.e. the probability of a feature occurring in in the positive label?

$$P(feature|positive) = ?$$

74

P(feature | label)

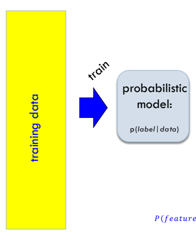
$$P(heads) = \frac{number\ of\ times\ heads\ came\ up}{total\ number\ of\ coin\ tosses}$$

Can we do the same thing here? What is the probability of a feature given positive, i.e. the probability of a feature occurring in in the positive label?

$$P(feature|positive) = \frac{number\ of\ positive\ examples\ with\ that\ feature}{total\ number\ of\ positive\ examples}$$

75

Training Naïve Bayes



1. Count how many examples have each label
2. For all examples with a particular label, count how many times each feature occurs
3. Calculate the conditional probabilities of each feature for all labels:

$$P(feature|label) = \frac{number\ of\ ``label''examples\ with\ that\ feature}{total\ number\ of\ examples\ with\ that\ label}$$

76

Classifying with Naïve Bayes

For each label, calculate the product of p(feature|label) for each label

yellow, curved, no leaf, 6oz

P(yellow|banana)*…*P(6oz|banana)

P(yellow|apple)*…*P(6oz|apple)

MAX

77

Naïve Bayes Text Classification

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that I loved it	I loved that I hated it

Given examples of text in different categories, learn to predict the category of new examples

Sentiment classification: given positive/negative examples of text (sentences), learn to predict whether new text is positive/negative

78

Text classification training

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that I loved it	I loved that I hated it

We'll assume words just occur once in any given sentence

79

Text classification training

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that loved it	I loved that hated it

We'll assume words just occur once in any given sentence

80

Training the model

Positive

I loved it
I loved that movie
I hated that loved it

Negative

I hated it
I hated that movie
I loved that hated it

For each word and each label, learn:

p(word | label)

81

Training the model

Positive

I loved it
I loved that movie
I hated that loved it

Negative

I hated it
I hated that movie
I loved that hated it

P(I | positive) = ?

$$P(word|label) = \frac{\text{number of times word occured in "label" examples}}{\text{total number of examples with that label}}$$

82

Training the model

Positive

I loved it
I loved that movie
I hated that loved it

Negative

I hated it
I hated that movie
I loved that hated it

P(I | positive) = 3/3 = 1.0

$$P(word|label) = \frac{\text{number of times word occured in "label" examples}}{\text{total number of examples with that label}}$$

83

Training the model

Positive

I loved it
I loved that movie
I hated that loved it

Negative

I hated it
I hated that movie
I loved that hated it

P(I | positive) = 1.0
P(loved | positive) = ?

$$P(word|label) = \frac{\text{number of times word occured in "label" examples}}{\text{total number of examples with that label}}$$

84

Training the model

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that loved it	I loved that hated it

P(I | positive) = 1.0
P(loved | positive) = 3/3

$$P(word|label) = \frac{number\ of\ times\ word\ occured\ in\ \text{"}label\text{"}\ examples}{total\ number\ of\ examples\ with\ that\ label}$$

85

Training the model

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that loved it	I loved that hated it

P(I | positive) = 1.0
P(loved | positive) = 3/3
P(hated | positive) = ?

$$P(word|label) = \frac{number\ of\ times\ word\ occured\ in\ \text{"}label\text{"}\ examples}{total\ number\ of\ examples\ with\ that\ label}$$

86

Training the model

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that loved it	I loved that hated it

P(I | positive) = 1.0
P(loved | positive) = 2/3
P(hated | positive) = 1/3
…

P(I | negative) = ?

$$P(word|label) = \frac{number\ of\ times\ word\ occured\ in\ \text{"}label\text{"}\ examples}{total\ number\ of\ examples\ with\ that\ label}$$

87

Training the model

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that loved it	I loved that hated it

P(I | positive) = 1.0
P(loved | positive) = 2/3
P(hated | positive) = 1/3
…

P(I | negative) = 1.0

$$P(word|label) = \frac{number\ of\ times\ word\ occured\ in\ \text{"}label\text{"}\ examples}{total\ number\ of\ examples\ with\ that\ label}$$

88

Training the model

Positive

I loved it

I loved that movie

I hated that loved it

Negative

I hated it

I hated that movie

I loved that hated it

P(I | positive) = 1.0
P(loved | positive) = 2/3
P(hated | positive) = 1/3
…

P(I | negative) = 1.0
P(movie | negative) = ?

$$P(word|label) = \frac{number\ of\ times\ word\ occured\ in\ "label"\ examples}{total\ number\ of\ examples\ with\ that\ label}$$

89

Training the model

Positive

I loved it

I loved that movie

I hated that loved it

Negative

I hated it

I hated that movie

I loved that hated it

P(I | positive) = 1.0
P(loved | positive) = 2/3
P(hated | positive) = 1/3
…

P(I | negative) = 1.0
P(movie | negative) = 1/3
…

$$P(word|label) = \frac{number\ of\ times\ word\ occured\ in\ "label"\ examples}{total\ number\ of\ examples\ with\ that\ label}$$

90

Classifying

P(I | positive) = 1.0
P(loved | positive) = 1.0
p(it | positive) = 2/3
p(that | positive) = 2/3
p(movie|positive) = 1/3
P(hated | positive) = 1/3

P(I | negative) = 1.0
p(hated | negative) = 1.0
p(that | negative) = 2/3
P(movie | negative) = 1/3
p(it | negative) = 2/3
p(loved | negative) = 1/3

Notice that each label has its own probability distribution

P(loved| positive)
P(loved | positive) = 2/3
P(no loved|positive) = 1/3

91

Trained model

P(I | positive) = 1.0
P(loved | positive) = 2/3
p(it | positive) = 2/3
p(that | positive) = 2/3
p(movie|positive) = 1/3
P(hated | positive) = 1/3

P(I | negative) = 1.0
p(hated | negative) = 1.0
p(that | negative) = 2/3
P(movie | negative) = 1/3
p(it | negative) = 2/3
p(loved | negative) = 1/3

How would we classify: "I hated movie"?

92

Trained model

P(I | positive) = 1.0
P(loved | positive) = 2/3
p(it | positive) = 2/3
p(that | positive) = 2/3
p(movie|positive) = 1/3
P(hated | positive) = 1/3

P(I | negative) = 1.0
p(hated | negative) = 1.0
p(that | negative) = 2/3
P(movie | negative) = 1/3
p(it | negative) = 2/3
p(loved | negative) = 1/3

P(I | positive) * P(hated | positive) * P(movie | positive) = 1.0 * 1/3 * 1/3 = 1/9

P(I | negative) * P(hated | negative) * P(movie | negative) = 1.0 * 1.0 * 1/3 = 1/3

93

Trained model

P(I | positive) = 1.0
P(loved | positive) = 2/3
p(it | positive) = 2/3
p(that | positive) = 2/3
p(movie|positive) = 1/3
P(hated | positive) = 1/3

P(I | negative) = 1.0
p(hated | negative) = 1.0
p(that | negative) = 2/3
P(movie | negative) = 1/3
p(it | negative) = 2/3
p(loved | negative) = 1/3

How would we classify: "I hated the movie"?

94

Trained model

P(I | positive) = 1.0
P(loved | positive) = 2/3
p(it | positive) = 2/3
p(that | positive) = 2/3
p(movie|positive) = 1/3
P(hated | positive) = 1/3

P(I | negative) = 1.0
p(hated | negative) = 1.0
p(that | negative) = 2/3
P(movie | negative) = 1/3
p(it | negative) = 2/3
p(loved | negative) = 1/3

P(I | positive) * P(hated | positive) * P(the | positive) * P(movie | positive) =

P(I | negative) * P(hated | negative) * P(the | negative) * P(movie | negative) =

95

Trained model

P(I | positive) = 1.0
P(loved | positive) = 2/3
p(it | positive) = 2/3
p(that | positive) = 2/3
p(movie|positive) = 1/3
P(hated | positive) = 1/3

P(I | negative) = 1.0
p(hated | negative) = 1.0
p(that | negative) = 2/3
P(movie | negative) = 1/3
p(it | negative) = 2/3
p(loved | negative) = 1/3

P(I | positive) * P(hated | positive) * P(the | positive) * P(movie | positive) =

P(I | negative) * P(hated | negative) * P(the | negative) * P(movie | negative) =

What are these?

96

Trained model

P(I | positive) = 1.0
P(loved | positive) = 2/3
p(it | positive) = 2/3
p(that | positive) = 2/3
p(movie|positive) = 1/3
P(hated | positive) = 1/3

P(I | negative) = 1.0
p(hated | negative) = 1.0
p(that | negative) = 2/3
P(movie | negative) = 1/3
p(it | negative) = 2/3
p(loved | negative) = 1/3

P(I | positive) * P(hated | positive) * P(the | positive) * P(movie | positive) =

P(I | negative) * P(hated | negative) * P(the | negative) * P(movie | negative) =

0. Is this a problem?

97

Trained model

P(I | positive) = 1.0
P(loved | positive) = 2/3
p(it | positive) = 2/3
p(that | positive) = 2/3
p(movie|positive) = 1/3
P(hated | positive) = 1/3

P(I | negative) = 1.0
p(hated | negative) = 1.0
p(that | negative) = 2/3
P(movie | negative) = 1/3
p(it | negative) = 2/3
p(loved | negative) = 1/3

P(I | positive) * P(hated | positive) * P(the | positive) * P(movie | positive) =

P(I | negative) * P(hated | negative) * P(the | negative) * P(movie | negative) =

Yes. They make the entire product go to 0 !

98

Trained model

P(I | positive) = 1.0
P(loved | positive) = 2/3
p(it | positive) = 2/3
p(that | positive) = 2/3
p(movie|positive) = 1/3
P(hated | positive) = 1/3

P(I | negative) = 1.0
p(hated | negative) = 1.0
p(that | negative) = 2/3
P(movie | negative) = 1/3
p(it | negative) = 2/3
p(loved | negative) = 1/3

P(I | positive) * P(hated | positive) * P(the | positive) * P(movie | positive) =

P(I | negative) * P(hated | negative) * P(the | negative) * P(movie | negative) =

Our solution: assume any unseen word has a small, fixed probability, e.g. in this example 1/10

99

Trained model

P(I | positive) = 1.0
P(loved | positive) = 2/3
p(it | positive) = 2/3
p(that | positive) = 2/3
p(movie|positive) = 1/3
P(hated | positive) = 1/3

P(I | negative) = 1.0
p(hated | negative) = 1.0
p(that | negative) = 2/3
P(movie | negative) = 1/3
p(it | negative) = 2/3
p(loved | negative) = 1/3

P(I | positive) * P(hated | positive) * P(the | positive) * P(movie | positive) = 1/90

P(I | negative) * P(hated | negative) * P(the | negative) * P(movie | negative) = 1/30

Our solution: assume any unseen word has a small, fixed probability, e.g. in this example 1/10

100

Full disclaimer

I've fudged a few things on the Naïve Bayes model for simplicity

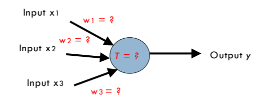
Our approach is very close, but it takes a few liberties that aren't technically correct, but it will work just fine ☺

If you're curious, I'd be happy to talk to you offline

101

Training neural networks

x_1	x_2	x_3	x_1 and x_2
0	0	0	1
0	1	0	0
1	0	0	1
1	1	0	0
0	0	1	1
0	1	1	1
1	0	1	1
1	1	1	0



1. start with some initial weights and thresholds
2. show examples repeatedly to NN
3. update weights/thresholds by comparing NN output to actual output

102