

ECE365: Introduction to NLP

Spring 2021

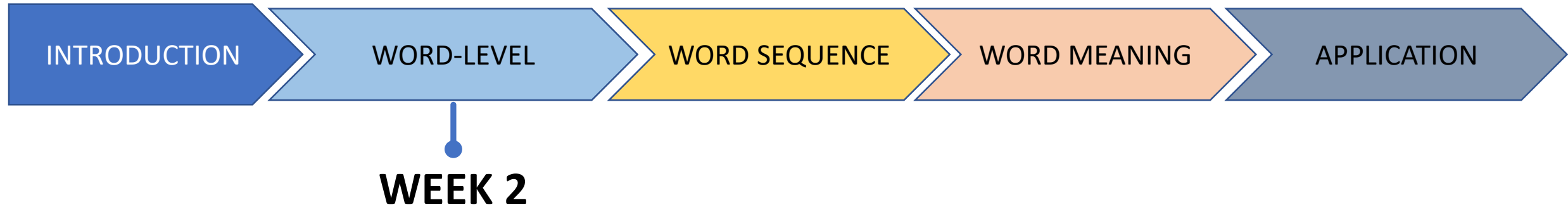
Lecture 3: Naïve Bayes Text Classification

[Reading J&M 4.1, 4.2, 4.3, 4.7]

Logistics

- Lab 2 posted; due
- Quiz on Tuesday 04/20
- Get on CBTF schedule if not already done
- ZJUI will have in-person exam

Course Progress



Natural language processing

- Extracting information from text (natural language understanding)
- Generating natural-language-like text automatically

What is the nature of natural language understanding we can perform using word-level information?

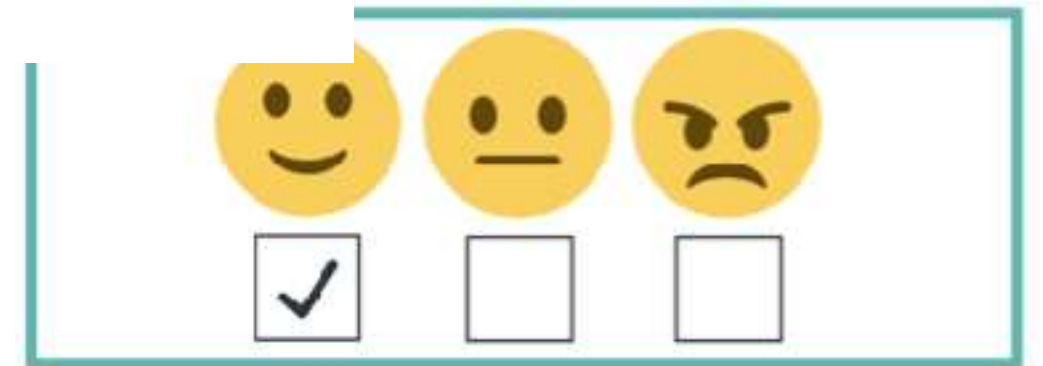
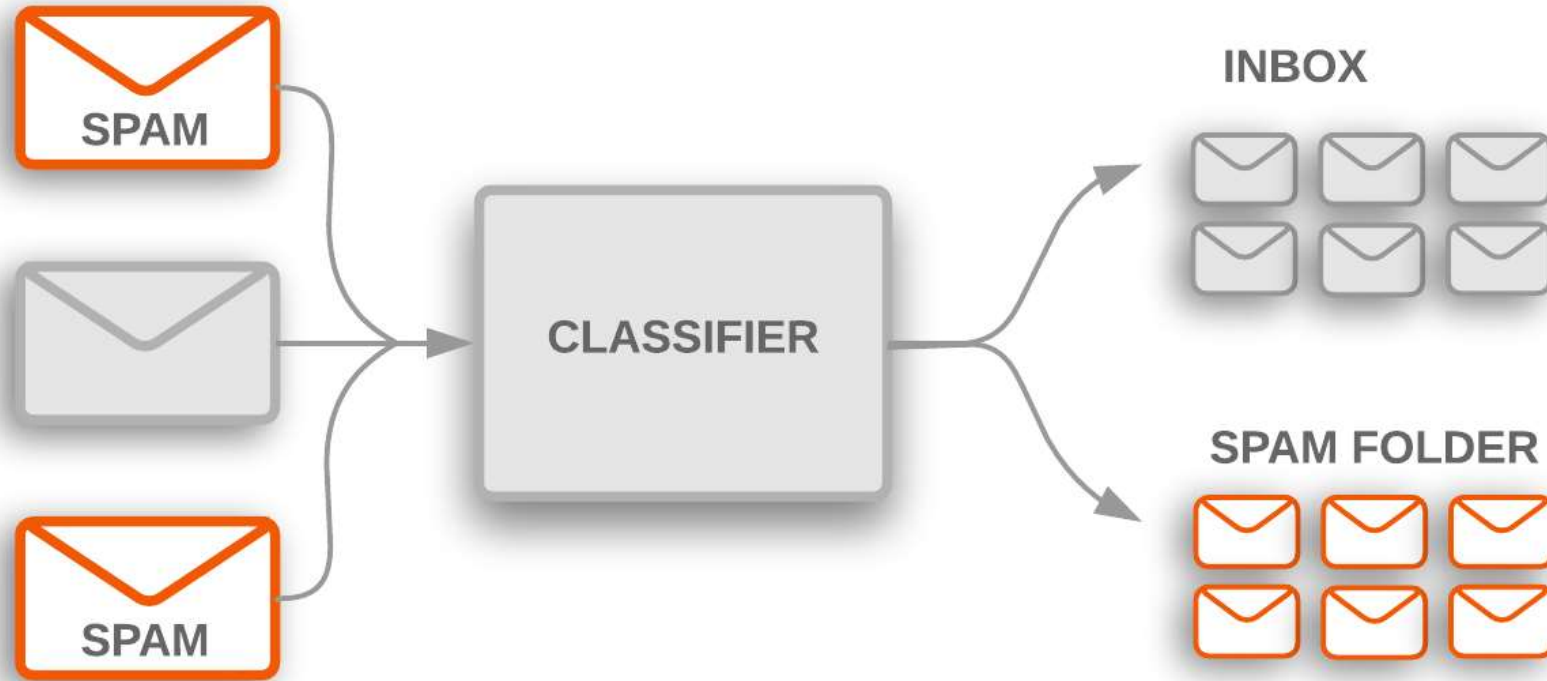
Text classification - 1

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam and the northern region, Tongking, was also a separate protectorate with its capital at Hanoi. The Annamese emperor, Khai Dinh, in theory ruled the two northern regions from Hue with the benefit of French protection, while Cochin-China was governed directly from Paris but in effect all three territories were ruled as colonies.
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets and even after years of comparative security she was still prepared for, still half expecting the old gibes to be revived. But whenever she was introduced, nothing greeted the amazing, all-revealing Clara but cries of “How delightful, how charming, how unusual, how fortunate,” and she could foresee a time when friends would name their babies after her and refer back to her with pride as the original from which inspiration had first been drawn.

Text Classification - 2

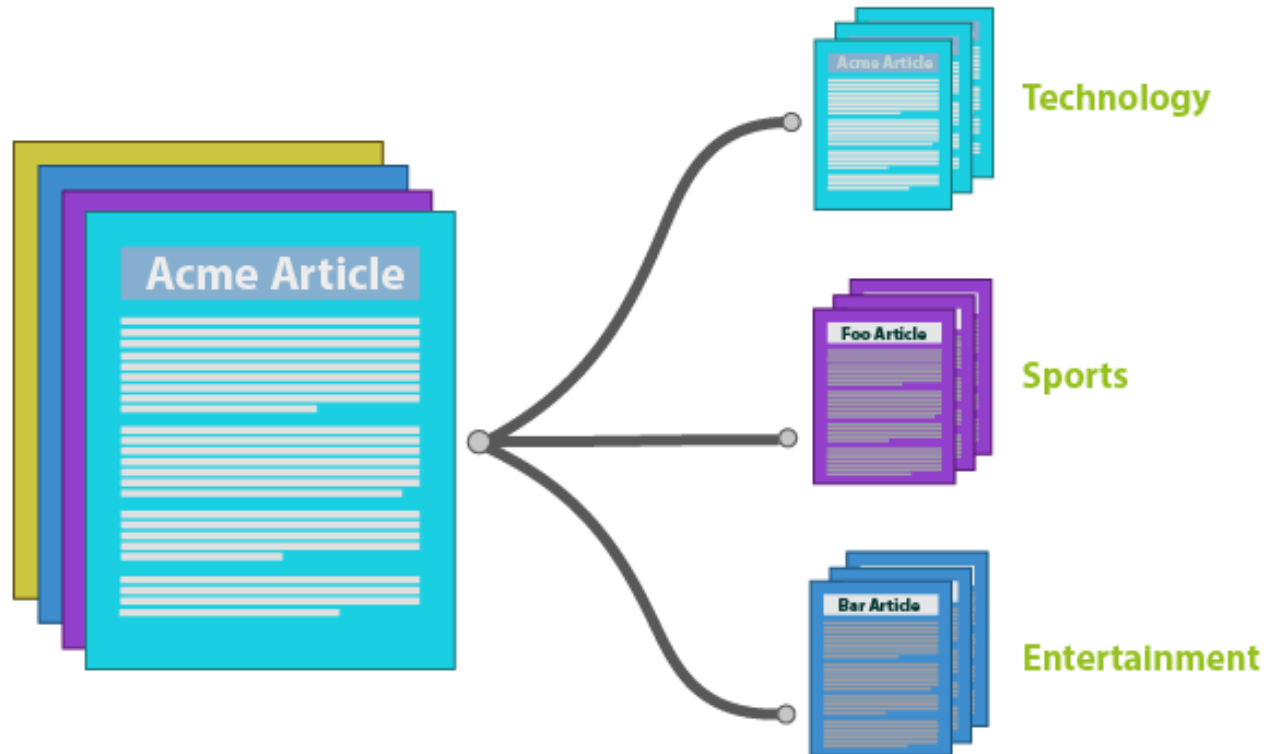
1. "The Necklace", by Guy de Maupassant, is a story about a woman named Ms. Loisel. She was petty and charming. She married a common little clerk. Her life was in poverty and depression. One day, She and her husband received an invitation to attend a ball. To conceal her impoverished family situations and show off her charm and beauty, she borrowed a necklace from her friend. Then, at the ball she did have a wonderful time. However, she lost the necklace after the ball.
2. Maupassant is one of the most influential writers in short fictions. He believes that "The writer's goal is to reproduce this illusion of life faithfully, using all the literary techniques at his disposal". In "The Necklace" Maupassant uses primarily symbolism to reveal his moral scheme that a person's preoccupation with appearance, materialistic existence, or idle pleasure is worthless and vain. By using the symbol of a Necklace, Maupassant is able to represent the vanity of Mathilde Loise, the main character, in a more visible way.

Text Classification – 3, 4



Text Classification-5

- Assigning subject categories, topics, or genres



The Task

- *Input:*
 - a document
 - a fixed set of classes
- *Output:* a predicted class

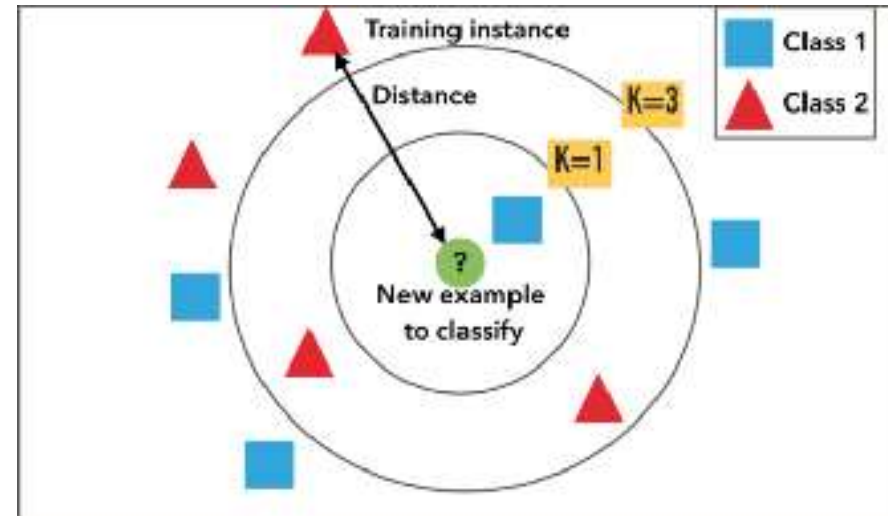
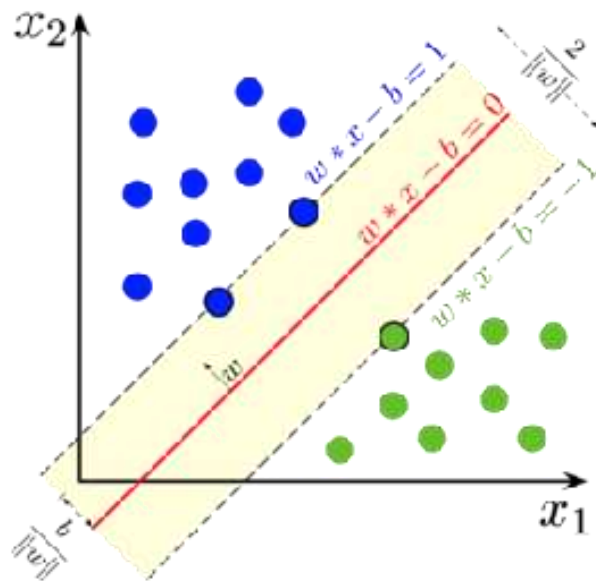
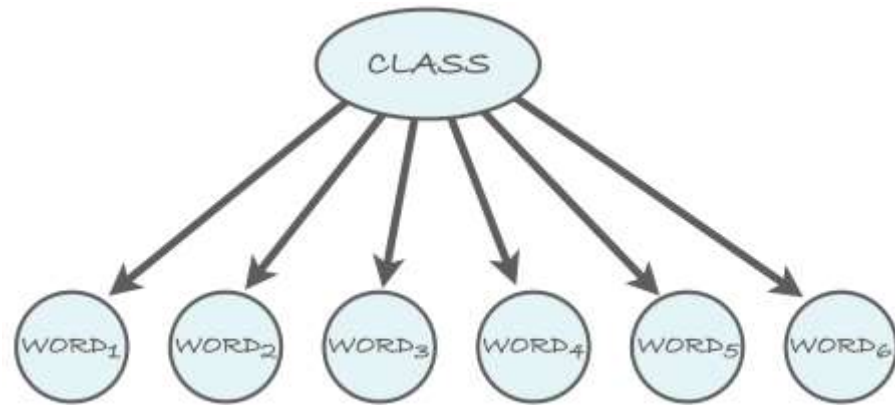
Rule-Based Classification

- Rules based on combinations of words or other features
 - spam: black-list-address ([*ithelpdesk.com, makemoney.com, spinthewheel.com, ...*]_[SEP]OR (“*dollars*” AND “*have been selected*”))
- Advantage: Accuracy can be high
- Disadvantage:

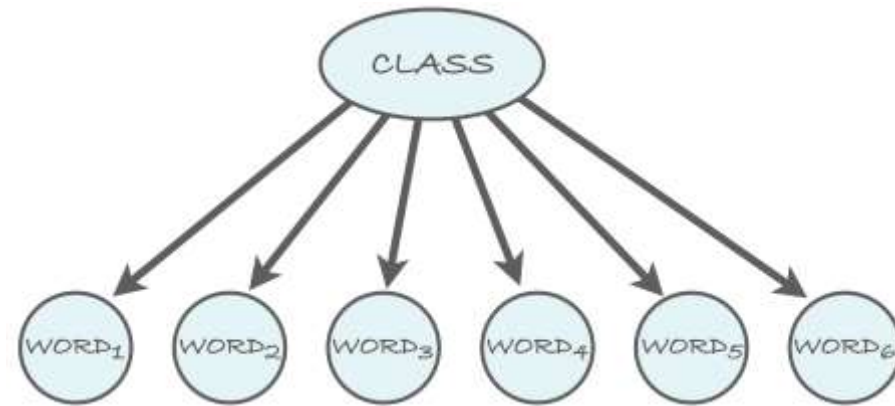
Supervised Learning: Let's use statistics!

- Data-driven approach
 - Let machine figure-out best patterns
- Inputs:
 - Set of J classes $C = \{c_1, c_2, \dots, c_J\}$
 - Set of n 'labeled' documents: $\{(d_1, t_1), \dots, (d_n, t_m)\}$
- Output
 - Trained classifier, $F : d \rightarrow c$

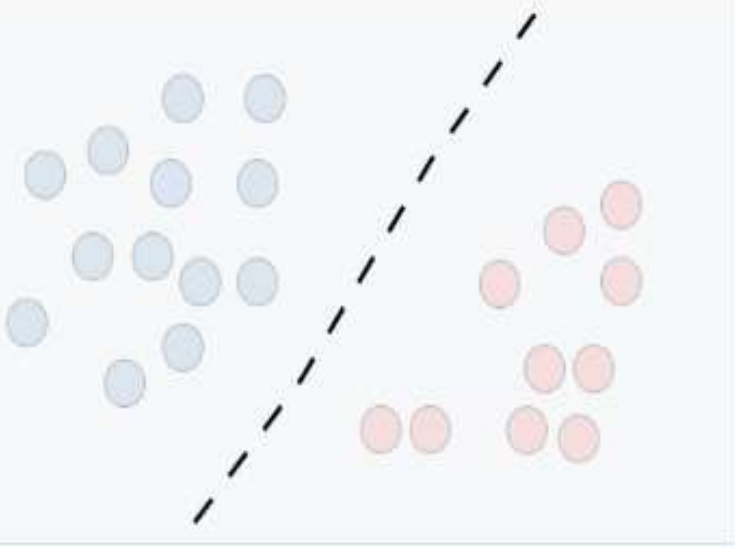
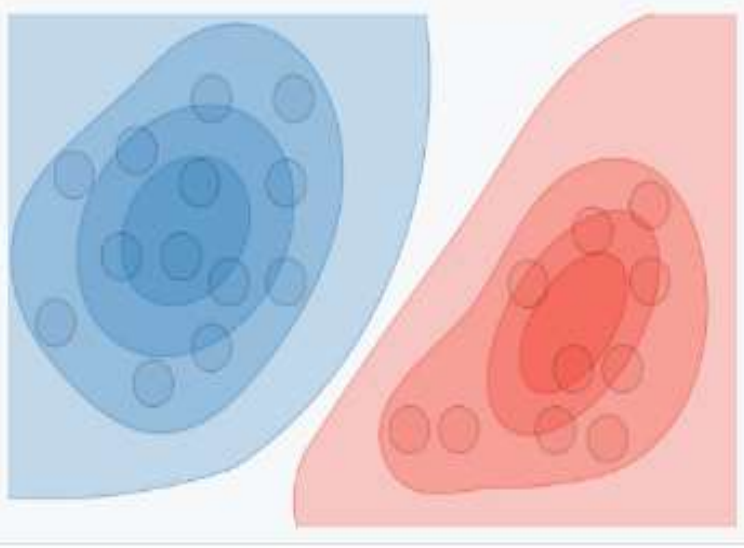
Types of Supervised Classifiers



Naïve Bayes Classifier



Naïve Bayes Classifier

	Discriminative model	Generative model
Goal	Directly estimate $P(y x)$	Estimate $P(x y)$ to then deduce $P(y x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration	 A scatter plot with blue and red circular data points. A dashed diagonal line separates the space into two regions, representing the decision boundary learned by a discriminative model.	 A scatter plot with blue and red circular data points. The background is filled with two overlapping, semi-transparent probability density contours: a blue one for the left cluster and a red one for the right cluster, representing the generative model's learned distributions.
Examples	Regressions, SVMs	GDA, Naive Bayes

Naïve Bayes Intuition

- Simple (“naïve”) classification method based on Bayes rule
- Bayes Rule:

Predicting the Class

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(c | d)$$

How to Represent $P(d|c)$?

- Option 1: Represent the entire sequence of words
compute

- Option 2: Bag of words
 - Assume position of word is irrelevant

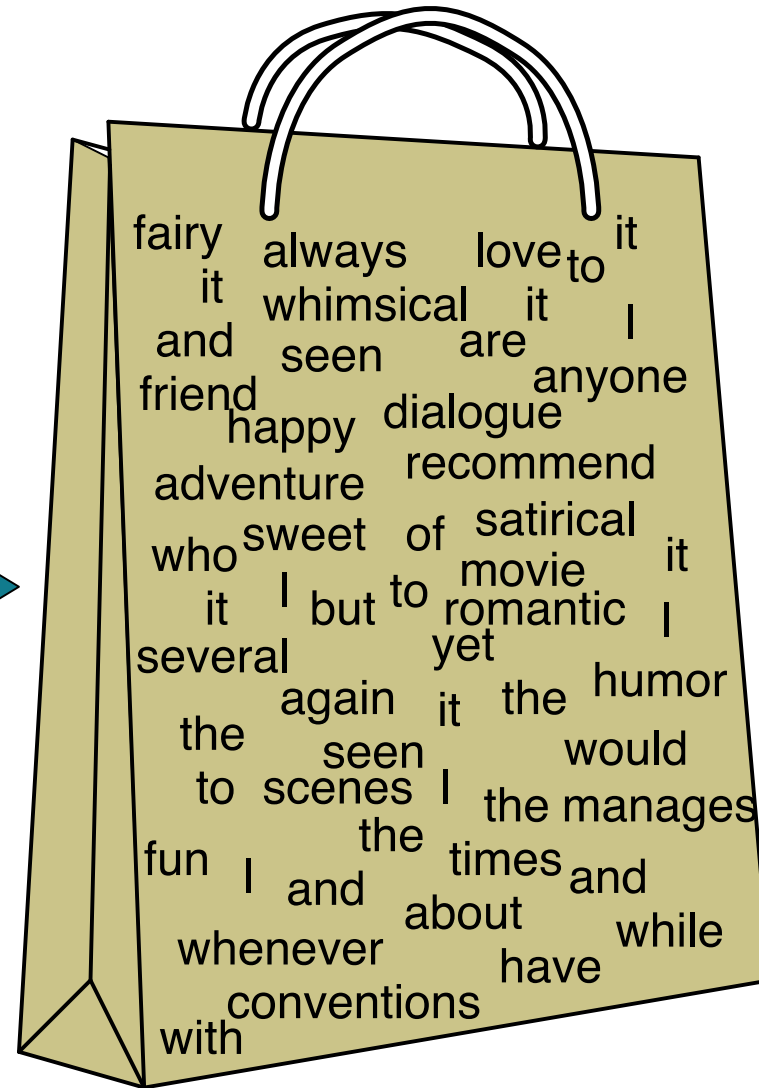
- Naïve part

features are

given class

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Predicting with Naïve Bayes

- We have: $c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$

Learning the Model

- Maximum likelihood estimates of probabilities
- Prior

Learning the Model

- Maximum likelihood estimates of probabilities
- Create mega-document for class c by concatenating all docs in this class (labels known)

Example

	Doc	Words	Class
Training	1	awesome game	sports
	2	closed shop	Not_sports
	3	close match	sports
	4	great shop	Not_sports
	5	awesome but forgettable game	sports

Test 6 great game ?

$V = \{\text{awesome, game, shop, close, match, great, forgettable, but}\}$

$|V| = K$

	Doc	Words	Class
Training	1	awesome game	sports
	2	closed shop	Not_sports
	3	close match	sports
	4	great shop	Not_sports
	5	awesome but forgettable game	sports

Test 6 great game ?

$V = \{\text{awesome, game, shop, close, match, great, forgettable, but}\}$

$$\hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}$$

$$\hat{P}(\text{great} | S) = 0?$$

Solution: Smoothing

- Laplace (add- α) smoothing
- $\hat{P}(w_i|c) =$

Putting it all together

Input a set of annotated documents $\{(d^{(i)}, c^{(i)})\} \ i = 1, \dots, n$

1. Compute vocabulary V of all words in training data

2. Calculate $\hat{P}(c_j) = \frac{\text{count}(c_j)}{\text{Total docs}}$

3. Calculate $\hat{P}(w_i | c_j) = \frac{(\text{count}(w_i, c_j) + 1)}{(\sum \text{count}(w_i, c_j)) + |V|}$, for $w_i \in V$

• (Prediction) Given document $d = \{w_1, w_2, \dots, w_K\}$

$$\hat{P}(c_j)$$

$$P(S) = 3/5$$

$$P(NS) = 2/5$$

$V = \{\text{awesome, game, shop, close, match, great, forgettable, but}\}$
 $|V| = 8$

Example

	Doc	Words	Class
Train	1	awesome game	S
	2	closed shop	NS
	3	close match	S
	4	great shop	NS
	5	awesome but forgettable game	S

$$\hat{P}(w_i|c_j) = \frac{(\text{count}(w_i, c_j) + 1)}{(\sum \text{count}(w_i, c_j) + |V|)}, \text{ for } w_i \in V$$

Test

awesome shop

?

$$P(\text{awesome} | S) = 2+1/8+8=3/16$$

$$P(\text{game} | S) = 2+1/8+8=3/16$$

$$P(\text{shop} | S) = 0+1/8+8=1/16$$

$$P(\text{close} | S) = 1+1/8+8=2/16$$

$$P(\text{match} | S) = 1+1/8+8=2/16$$

$$P(\text{great} | S) = 0+1/8+8=1/16$$

$$P(\text{forgettable} | S) = 1+1/8+8=2/16$$

$$P(\text{awesome} | NS) = 0+1/4+8=1/12$$

$$P(\text{game} | NS) = 0+1/4+8=1/12$$

$$P(\text{shop} | NS) = 2+1/4+8=3/12$$

$$P(\text{close} | NS) = 1+1/4+8=2/12$$

$$P(\text{match} | NS) = 0+1/4+8=1/12$$

$$P(\text{great} | NS) = 1+1/4+8=2/12$$

$$P(\text{forgettable} | NS) = 0+1/4+8=1/12$$

$$P(S/\text{test}) = 3/16 * 1/16 \\ = 0.011$$

$$P(NS/\text{test}) = 1/12 * 3/12 \\ = 0.021$$

Practical Issues

- Not necessarily only words
 - URLs, email addresses, Capitalization, ...
- Domain knowledge very useful
- Use log scale operations instead of multiplying probabilities
 - To prevent underflow of floating point due to low probabilities

Advantages of NB

- Very fast, low storage requirements
- Robust to irrelevant features
- Very good in domains with equally important features
- Optimal if independence assumptions hold
- A good dependable baseline classifier for text classification

Evaluation

Binary Classification

Actual

Predicted

	Positive	Negative
positive	100	5
negative	45	100

Evaluation

Actual

Predicted

	Positive	Negative
positive	100	5
negative	45	100

True positive: Predicted + and actual +

True negative: Predicted - and actual -

False positive: Predicted + and actual -

False negative: Predicted - and actual +

Precision and Recall

- Precision: % of selected classes that are correct

$$\text{Precision} = \frac{TP}{TP+FP}$$

- Recall: % of correct items selected

$$\text{Recall} = \frac{TP}{TP+FN}$$

F-Score

- Harmonic mean of Precision and Recall

$$F_1 = \frac{2PR}{P+R}$$

Summary