2024-2 Unstructured Data Analysis
2024.11.06

Document Classification

Week 10 - Code Practice

Professor: Misuk Kim
Teaching Assistant: Minjoo Son
minjoo77@hanyang.ac.kr

HANYANG UNIVERSITY

Contents

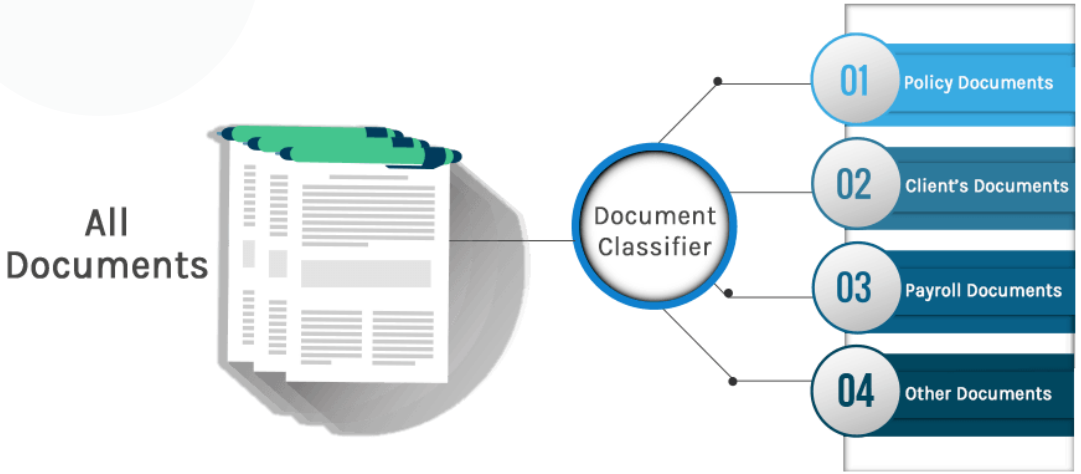
1. Introduction

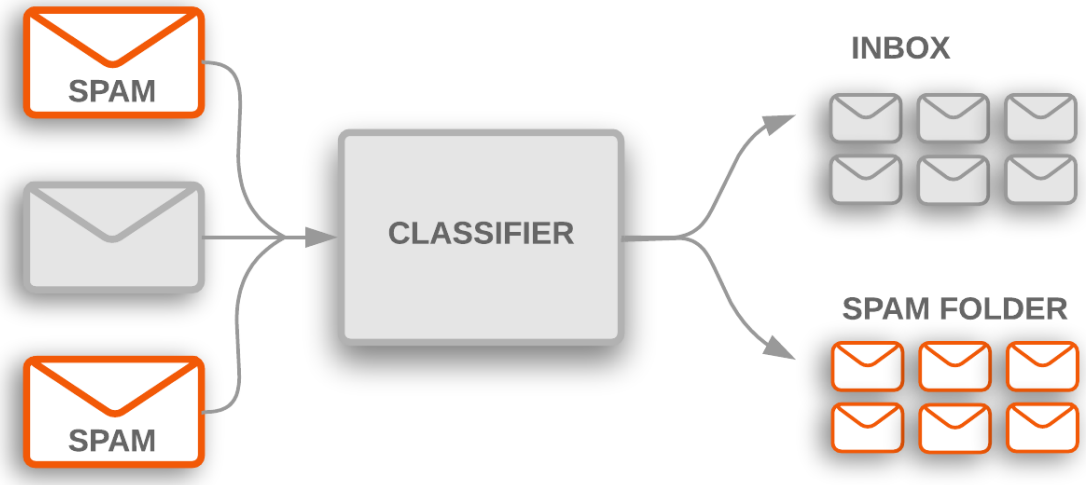
❖ Week 10 Objective

- Document Classification (Vector Space Model)
  - Naive Bayes Classifier (in-Class)
  - k-Nearest Neighbor Classifier (Assignment)

❖ Document Classification

- Document classification generally refers to the task of categorizing a given document into predefined classes.
  - News article categorization, Spam email classification





2. Document Classification

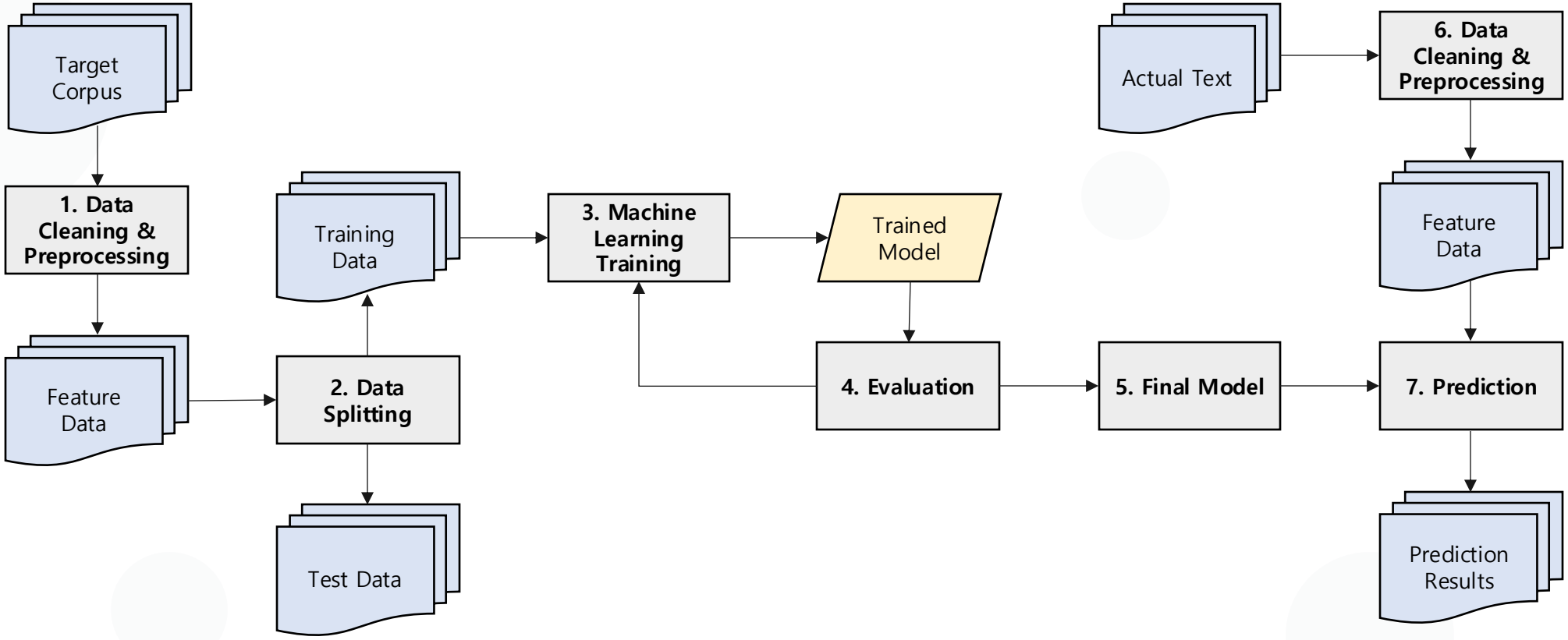
❖ Machine learning approach

- Various machine learning methods can be used for document classification, including Naive Bayes, k-Nearest Neighbors, logistic regression, decision trees, SVM, etc.
- Among these, Naive Bayes is particularly widely used, to the extent that it is specialized for text classification.
- For training, it is necessary for all documents or texts to have labels or predefined classes.

	ArticleId	Text	Category	CategoryId
0	1833	worldcom ex bos launch defence lawyer defendin...	business	0
1	154	german business confidence slide german busine...	business	0
2	1101	bbc poll indicates economic gloom citizen majo...	business	0
3	1976	lifestyle governs mobile choice faster better ...	tech	1
4	917	enron boss 168m payout eighteen former enron d...	business	0
...	...	...	...	...
1485	857	double eviction big brother model caprice holb...	entertainment	4
1486	325	dj double act revamp chart show dj duo jk joel...	entertainment	4
1487	1590	weak dollar hit reuters revenue medium group r...	business	0
1488	1587	apple ipod family expands market apple expande...	tech	1
1489	538	santy worm make unwelcome visit thousand websi...	tech	1

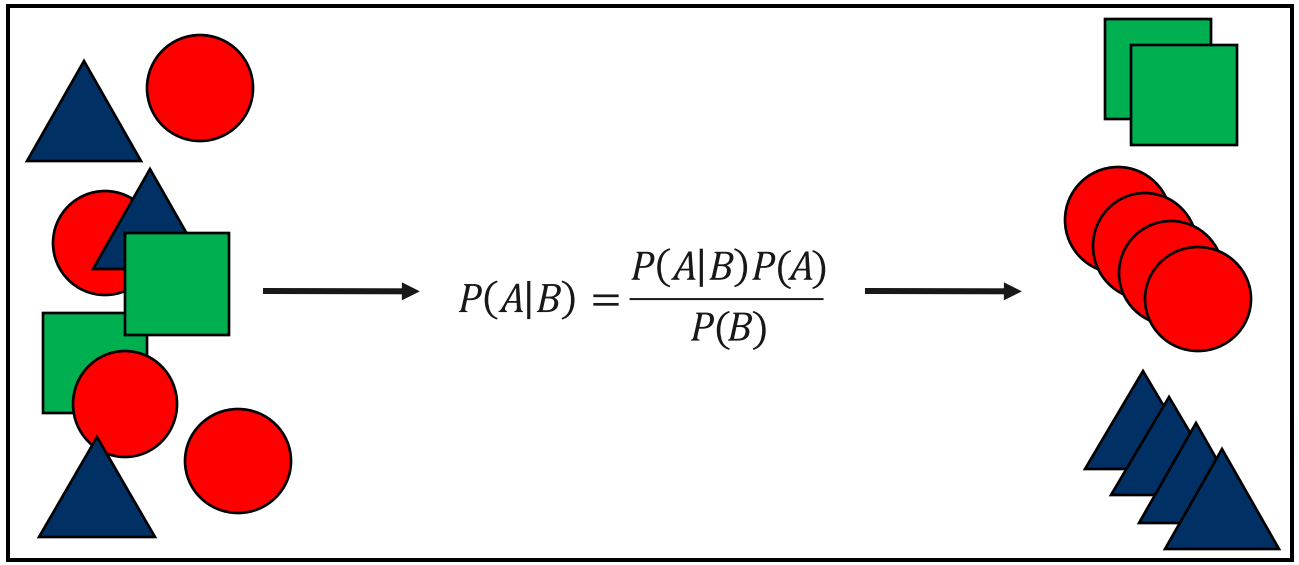
1490 rows × 4 columns

❖ Understanding Machine Learning and Document Classification Processes



1. Naive Bayes Classifier

- It is a probability-based classification algorithm that relies on the assumption of conditional independence of the features of the given data.
- This algorithm uses Bayes' Theorem to calculate the probability of belonging to a specific class and then assigns the class with the highest probability.
- The term "Naive" is used because it assumes that each feature is independent.



1. Naive Bayes Classifier

- Scikit-learn provides classes for Naive Bayes in 'sklearn.naive_bayes'
  - https://scikit-learn.org/1.5/modules/naive_bayes.html

Classifier	Concept	Use for Text Data	
		Recommended Situations	Example
Gaussian Naive Bayes	It assumes that each feature follows a Gaussian distribution (normal distribution), making it suitable for handling continuous data.	It is suitable for use with data that has continuous features.	When specific attributes of a document (e.g., length, word count, etc.) are continuous.
Multinomial Naive Bayes	It is primarily used for text classification and assumes that each feature follows a multinomial distribution.	It is suitable for classification based on word frequency or occurrence counts in text documents.	news articles and email spam filtering.
Complement Naive Bayes	A variation of Multinomial Naive Bayes, it is designed to address the class imbalance problem.	It is useful when the dataset is imbalanced (when one class significantly outnumbers the others).	When classification is required from a large number of categories into a few categories.
Bernoulli Naive Bayes	It is suitable for handling binary feature (i.e., presence/absence) data.	It is used when considering the presence or absence of words within a document (binary features).	It is useful when the dataset is imbalanced (when one class significantly outnumbers the others).
Categorical Naive Bayes	It is used when each feature has categorical data.	-	-

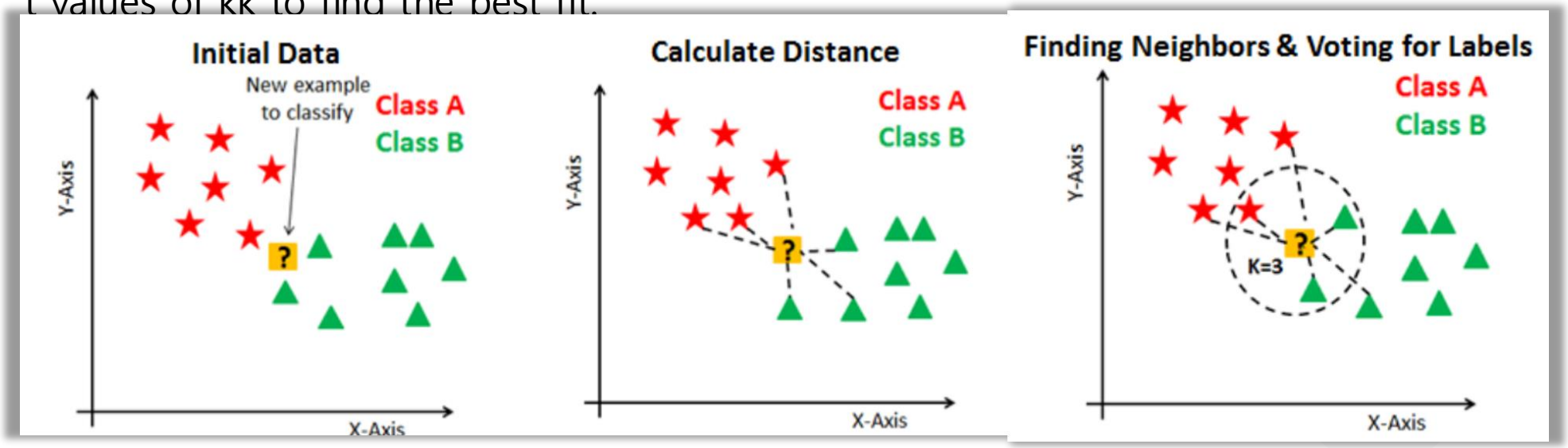
2. Document Classification

1. Naive Bayes Classifier

- Code Practice
  - Code is available on GitHub: https://github.com/ming9oori/Unstructured-Data-Analysis

2. k-Nearest Neighbor

- Classification is done by assigning a new datapoint to the class of the k nearest neighbors.
  - If k=1, the class of the new datapoint is assigned based on the class of the single nearest neighb or.
  - If k=5, the class is assigned based on the majority class among the 5 nearest neighbors.
  - Typically, when there is an even number of classes, kk is chosen as an odd number to avoid ties.
  - The optimal value of kk varies by dataset, so it is recommended to test performance with differen t values of kk to find the best fit.



2. k-Nearest Neighbor

- Scikit-learn provides classes for KNN classifiers in sklearn.neighbors.KNeighborsClassifier
  - https://scikit-learn.org/dev/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

❖ Assignment 3 – Document Classification

- Using the code from "Week 10 – Document Classification.ipynb," carry out document classification using Scikit-Learn's KNN classifier.
- Refer to the Scikit-Learn documentation on the KNN classifier's parameters and adjust them as needed.
  - https://scikit-learn.org/dev/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- Save the file as 'Assignment_YourName_YourStudentID.ipynb' and submit it to the 'Code Practice Assignment 3 – Document Classification' section under Assignments
- Due Date: 23:00 on Monday, November 11th.

Q & A

Thank you for your attention. Any questions are welcome!

Minjoo Son