Homework

August 20, 2025

1 Objectives

By the end of this assignment you should be able to:

1. Load and explore a real-world text dataset.
   (a) In particular from spam and ham from the spam.csv
2. We provide an example on how to
   (a) Clean and preprocess textual data.
   (b) Convert text into a numerical feature representation (Bag-of-Words, TF-IDF, n-grams).
   (c) Reduce the dimensionality to 100 vectors from sparse to dense matrix representation
3. Train a logistic regression classifier for binary classification
4. Report Results using ROC curves

2 Problem Statement

You will build a spam detector for short SMS messages.

1. Given a message, the model must decide whether it is spam or ham (legitimate).
2. The dataset contains 5 574 labeled SMS messages (4825 ham, 747 spam).
   (a) Here you have a problem of unbalanced data set, Any idea how to fix it?
3. Results a clean, well-documented notebook that trains, evaluates, and interprets a logistic regression reporting ROC curves