Machine Learning - Practice Problems 1

Universidad de La Sabana
Santiago Toledo-Cortés
Facultad de Ingeniería 2025-2

1. Kaggle Python Tutorial

Complete the tutorial Kaggle Python Tutorial on Machine Learning.

2. Probability and Linear Algebra

Let $D = \{d_1, \ldots, d_n\}$ be a set of documents and $T = \{t_1, \ldots, t_m\}$ a set of terms (words). Define:

- $TD = (TD_{i,j})_{i=1\ldots m, j=1\ldots n}$ as a matrix where $TD_{i,j}$ represents the frequency of term t_i in document d_j.
- l_i as the length (number of characters) of term t_i.
- $L = (l_1, \ldots, l_m)$ as a column vector of term lengths.
- A process where a document d_j is chosen uniformly at random, and a term t_i from d_j is chosen with probability proportional to its frequency in d_j.

For each of the following, provide:

1. A mathematical expression using TD, L, constants (scalars, vectors, matrices), and linear algebra operations.
2. A corresponding NumPy expression that evaluates to the requested matrix, vector, or scalar.
3. The computed result given:

$$TD = \begin{bmatrix} 2 & 3 & 0 & 3 & 7 \\ 0 & 5 & 5 & 0 & 3 \\ 5 & 0 & 7 & 3 & 3 \\ 3 & 1 & 0 & 9 & 9 \\ 0 & 0 & 7 & 1 & 3 \\ 6 & 9 & 4 & 6 & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 5 \\ 2 \\ 3 \\ 6 \\ 4 \\ 3 \end{bmatrix}$$

(a) Joint Probability Matrix $P(T, D)$

Each position $P(T, D)_{i,j}$ represents $P(t_i, d_j)$.

(b) Conditional Probability Matrix $P(T|D)$

(c) Conditional Probability Matrix $P(D|T)$

(d) Vector $P(D)$

(e) Vector $P(T)$

(f) Expected Value $E[l]$

Expected value of the length of a randomly chosen term.

(g) Variance Var(l)

Variance of the length of a randomly chosen term.

Note: The solutions should be presented in a Jupyter Notebook, in groups of 1 person. For the grade, you must present your work at the beginning of the next class.