CS4641 Machine Learning - Homework 3

Bo Dai

Deadline: 03/24 Mon, 23:59 PM

- Submit your answers as 1) one single PDF file to *HW3* and 2) one Jupyter Notebook file to *HW3_code* on Gradescope. **IMPORTANT: The solution to each problem/subproblem must be on a separate page. When submitting to Gradescope, please make sure to mark the page(s) corresponding to each problem/subproblem.**
- You will be allowed 2 total late days (48 hours) without penalty for the entire semester. Once those days are used, you will be penalized according to the following policy:
  - Homework is worth full credit before the due time.
  - It is worth 75% credit for the next 24 hours.
  - It is worth 50% credit for the second 24 hours.
  - It is worth zero credit after that.
- You are required to use Latex, or word processing software, to generate your solutions to the written questions. Handwritten solutions WILL NOT BE ACCEPTED.

1 Naive Bayes

*Please submit the solution to this problem to **HW3** on Gradescope.*

In medical diagnosis, doctors often need to determine whether a patient has a certain disease based on symptoms. Suppose a hospital research team has collected historical patient data with three key binary features: **Hypertension** (X_1), **High Cholesterol** (X_2), and **Family History** (X_3), along with the diagnosis result (Y) indicating whether the patient has the disease.

In this problem, we assume that the features X_1, X_2, X_3 are conditionally independent given Y.

The dataset is given as follows:

Patient ID	Hypertension (X_1)	High Cholesterol (X_2)	Family History (X_3)	Disease (Y)
1	1	1	1	1
2	1	0	1	1
3	0	1	0	0
4	1	1	0	0
5	0	0	1	0
6	1	0	0	0
7	0	1	0	1
8	0	0	1	0

A patient is considered to have the disease ($Y = 1$) if they exhibit symptoms matching past cases where the disease was diagnosed, and they are considered healthy ($Y = 0$) otherwise.

Using the dataset above, answer the following questions:

- **[5 points]** Compute the prior probabilities of having the disease and being healthy.
- **[15 points]** Compute the conditional probabilities of each symptom given the disease status. That is, calculate $P(X_j = 1 \mid Y = 1)$ and $P(X_j = 1 \mid Y = 0)$ for each $j = 1, 2, 3$ corresponding to Hypertension, High Cholesterol, and Family History.
- **[20 points]** A new patient arrives at the hospital with the following symptoms: having hypertension, not having high cholesterol, and having a family history of the disease. Using Bernoulli Naive Bayes, compute the posterior probabilities of this patient having the disease ($P(Y = 1 \mid X_1 = 1, X_2 = 0, X_3 = 1)$) and being healthy ($P(Y = 0 \mid X_1 = 1, X_2 = 0, X_3 = 1)$).

2 K-means

*Please submit the completed notebook to **HW3_code** on Gradescope.*

2.1 Description

For this problem, you will implement the K-means algorithm using the `Numpy` library. Concretely, you need to complete the TODOs in the following `.ipynb` file.

`https://drive.google.com/file/d/1dpxpet2-CdGT4v-8krH9o1ddp9vciXXa/view?usp=sharing`

2.2 Implementation TODOs

1. **Setup**
   - *No implementation required (code provided)*
2. **Helper Function Implementation[35 points]**
   - `euclidean_distance()`: 5 points
   - `initialize_centroids()`: 10 points
   - `assign_clusters()`: 10 points
   - `update_centroids()`: 10 points
3. **K-means Algorithm Implementation [25 points]**
   - Implement the main `k_means()` function using your helper functions
4. **Comparison with scikit-learn Implementation**
   - *No implementation required (code provided)*
5. **Elbow Method Analysis**
   - *No implementation required (code provided)*

2.3 Implementation and Submission Rules

- You can either download the notebook and complete the TODOs in your local environment, or make a copy and modify it on Google Colab.
- Use **only** `Numpy` for your implementation (no `scikit-learn`/`scipy` for clustering functions)
- Please submit your completed `.ipynb` file **including all the outputs of each block** to HW3_code on Gradescope.