Streams Processing Tutorials

May 15, 2023

1 Vectors and Matrices

1. *Pacients, symptoms, and treatments:* Assume you receive a dataset from a hospital with m patients, n symptoms, and p treatments. Your data is organized in two matrices $S \in \mathbb{R}^{m \times n}$, and $T \in \mathbb{R}^{m \times p}$ such that

$$S_{ij} = \begin{cases} 1 & \text{if patient } i \text{ has symptom } j \\ 0 & \text{otherwise} \end{cases}$$

$$T_{ik} = \begin{cases} 1 & \text{if patient } i \text{ was treated with } k \\ 0 & \text{otherwise.} \end{cases}$$

   (a) What is the meaning of the second column of S? And what about row 500 of T?

   (b) We define the vector of all ones as $\mathbf{1}$ and the transpose of a matrix A as A^T. Describe in plain English the following quantities, and, further, mention dimensions and entrywise expressions:

      i. $S\mathbf{1}$.
      ii. $S^T\mathbf{1}$.
      iii. S^TS.
      iv. SS^T.

   (c) Consider matrix $P \in \mathbb{R}^{n \times p}$ where P_{jk} is the total number of patients with symptom j that received treatment k. Express P in matrix notation as a function of matrices S and T.

   (d) How would your conclusions change if the encoding of the binary variables changed from $\{0, 1\}$ to $\{-1, 1\}$?

2. *Cycles in a graph:* Consider that your next assignment in the streams processing class is to rank a set of m images according to perceived safety, using only n pairwise comparisons that were crowdsourced in a website. The images are the nodes of a directed graph. Those nodes are connected by directed edges derived from the pairwise comparisons: if there was a comparison where image i was considered safer than j, then the edge between them will be incident to j. Your assignment for the class is to detect cycles on the graph. We will define a cycle of length ℓ as a path that starts in some node i and ends after ℓ hops again in node i. Consider a fragment of this graph depicted in Figure 1,

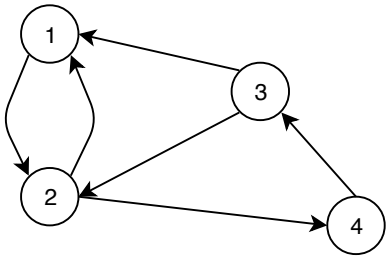

Figure 1 – Fragment of the graph of images and pairwise comparisons.

and the associated adjacency matrix

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

(a) Compute the number of cycles of length $\ell = 10$ and $\ell = 11$. (*Hint:* See what happens when you compute A^2).

(b) Say how many cycles of length $\ell = 11$ start and finish at each of the four nodes. Note that the sum of cycles for all nodes should be the same as the related value from the previous question.

2 Intro to learning from data streams

1. *Feasibility of learning:* Recall the problem of the marketing pollster from lecture 1 of the module. We sampled milk-buying customers of a supermerket to see how many bought brand A and how many bought brand B. We saved the result of our sampling in variables

$$z_i,\ i = 1, \cdots, n$$

where $z_i = 1$ if customer i bought brand A and $z_i = 0$ if customer i bought brand B. We computed the sample mean as

$$\nu = \frac{1}{n} \sum_{i=1}^{n} z_i.$$

The true mean is the probability of buying milk A for the overall population of milk buyers, which is μ.

   (a) Explain in your own words/math how this reasoning applies to generalization error, as briefly mentioned in class and further explained in the slides.

   (b) We know that the Hoeffding inequality

$$\mathbb{P}\left(|\nu - \mu| > \epsilon\right) \leq 2e^{-2\epsilon^2 n}$$

holds for any $\epsilon > 0$ when applied to our supermarket problem. Assume we computed $\nu = 0.9$ from a sample of size $n = 200$ and that we want to be within a tolerance $\epsilon = 0.01$.

      i. Can we say that μ will *never* be 0.1?

      ii. What is the Hoeffding inequality telling us about μ?

      iii. Would you recommend your boss to gather more data, or, in your opinion, would this dataset be enough?

2. *Temperature prediction:* You are given a dataset with two weeks of hourly temperature data from Lisbon, with $15{\cdot}24 = 360$ data points, $x(t)$. You are asked to design a predictor $\hat{x}(t+1)$ of Lisbon's temperature one hour ahead. The standard deviation of some predictors can be estimated via the root mean squared error (RMSE)

$$\text{RMSE}(\hat{x}(t)) = \sqrt{\frac{1}{N} \sum_{t=M+1}^{T} (\hat{x}(t) - x(t))^2},$$

where $N = T - M$. To answer the following questions load the file `temperatureLisbon.csv` in a colab notebook.

(a) What is the RMSE of the average estimator $\hat{x}_1(t+1) = \frac{1}{t}\sum_{\tau=1}^{t} x(\tau)$?

(b) What about $\hat{x}_2(t+1) = x(t)$? And $\hat{x}_3(t+1) = x(t-23)$?

(c) Consider an autorregressive (AR) model of lag M

$$\hat{x}_4(t+1) = a_1 x(t) + a_2 x(t-1) + \cdots + a_M x(t-M+1),$$

with $t = M, M+1, \cdots$, and a_n are the model parameters. It is possible to choose the model parameters using the observed dataset, by minimizing the sum of squares of the prediction errors $x(t) - \hat{x}(t)$.

i. Formulate the problem as a least squares problem

$$\underset{z}{\text{minimize}} \, \|Az - y\|^2,$$

and define matrix A, vectors z and y using the model parameters a_n, the data points $x(t)$, and the estimator $\hat{x}_4(t)$.

ii. Consider $M = 8$ and computationally solve the least squares problem you obtained in the previous question, plot the data and your prediction, and compute the RMSE for estimator $\hat{x}_4(t)$.

iii. *(Optional)* Prove analytically that the average of prediction errors for the estimator $\hat{x}_4(t)$ is zero.

3 Dimensionality reduction

1. *Dimensionality reduction:* You will be working with `data1.csv`.

   (a) Download `data1.csv` and perform some EDA. What can you say about the dataset? (*Suggestion:* look for the `describe` function, and the plotly plot engine).

   (b) What dimensionality reduction methods could apply to these data, and why?

   (c) Experiment with the methods you chose. Comment on the results. You can (and should) use standard libraries, but make sure your data meets the necessary conditions for the methods to be applicable.

4 Learning under Concept Drift

1. *Classification:* Follow this tutorial by Bravin Wasike. Apply the river implementation of the Hoeffding Adaptive Tree classifier to a categorical variable in your dataset.

2. *Regression:* Test drift-aware regression algorithms in river for continuous variables of your project dataset. If your task is classification use the dataset AirlinePassengers in the datasets subpackage of river.

3. *Concept Drift:* Apply ADWIN drift detector to your dataset. Analyse the results.

5 Incremental and online learning

1. *Online optimization and incremental learning:* Copy the notebook PS-Regression to your Google drive. Run it and experiment with different optimizers. In this notebook, what is the function `get_hour_and_weekday(x)` doing?

2. *Your data problem:* Refine your model to answer your research question. Evaluate your error. Does your model takes advantage of all the available information? Is there some feature transformation you could try?

3. *Going from batch to online learning. The example of the Standard-Scaler:* Because we obtain our data from a stream, many common operations from the batch ML pipeline have to be computed differently. Consider the trivial operation of scaling the data so they have mean zero and standard deviation one. To do so we simply have to subtract the mean of each feature to each value and then divide the result by the standard deviation of the feature. How can we compute the mean and standard deviation in a streaming setting? (hint: write the expression of the mean at $t + 1$, $\bar{x}_{t+1}$, in terms of $\bar{x}_t$.)

6 Ensemble methods and learning from imbalanced data