Supplementary Material for: A Dataset and Benchmark for Automatically Answering and Generating Machine Learning Final Exams

Sarah Zhang
EECS
MIT
sazhang@mit.edu

Reece Shuttleworth
EECS
MIT
rshuttle@mit.edu

Derek Austin
CS
Columbia University
da2986@columbia.edu

Yann Hicke
CS
Cornell University
ylh8@cornell.edu

Leonard Tang
Mathematics
Harvard University
leonardtang@college.harvard.edu

Sathwik Karnik
EECS
MIT
skarnik@mit.edu

Darnell Granberry
EECS
MIT
darnellg@mit.edu

Iddo Drori
EECS and CS
MIT and Columbia University
idrori@mit.edu, idrori@cs.columbia.edu

A Appendix

1. Submission introducing new datasets must include the following in the supplementary materials:
   (a) **Documentation and intended uses.** Documentation can be found in the README files of the GitHub repository https://github.com/idrori/mlfinalsQ. The authors intend for the dataset to be used by the machine learning and broader research community to improve
   (b) **URL to dataset.** The full dataset can be accessed and downloaded at https://github.com/idrori/mlfinalsQ.
   (c) **Author statement.** The authors bear all responsibility in case of violation of rights, etc., and confirm the data license.
   (d) **Hosting, licensing, and maintenance plan.** The data and code is hosted and maintained on GitHub under an MIT license.
2. To ensure accessibility, the supplementary materials for datasets must include the following:
   (a) **Links to access the dataset and its metadata.** The dataset and its metadata is accessible at https://github.com/idrori/mlfinalsQ.
   (b) **Reading the dataset.** Questions in the dataset are presented in json file format, with the following fields:

   The dataset is also available to download as a CSV file with the fields described in 1 as column headers.

Submitted to the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks. Do not distribute.

Field	Description
Semester	The semester the question's final was given in (ex. Fall 2017)
Question Number	The number of the question from the final (ex. 1, 2...)
Part Number	The question part label the question has (ex. a, b.i)
Points	The number of points the question is worth
Topic	The primary machine learning topic that the question targets Topics are regression, classifiers, logistic regression, features, loss functions, neural networks, CNNs, MDPs, RNNs, reinforcement learning, clustering, and decision trees
Type	Text if the question only relies on text, Image if the question relies on an image
Question	The original question text as presented from the source
Solution	The solution to the question

Table 1: Dataset json fields and their descriptions.

(c) **Long-term preservation.** We ensure the longevity of this dataset by keeping it publicly available in a GitHub repository.

(d) **Explicit license.** The code and data is licensed under the MIT license in the repository.

(e) Add structured metadata to a dataset's meta-data page using Web standards (like schema.org and DCAT): This allows it to be discovered and organized by anyone. If you use an existing data repository, this is often done automatically. `https://github.com/idrori/mlfinalsQ/blob/main/data/schema_dataset.json`

(f) **Dataset Identification.** Our data and code are maintained in a GitHub repository, allowing for easy access, Our dataset thus does not have a DOI.

3. **Reproducibility.** Data, code, and evaluation procedures for reproducing the benchmark results in this paper are available at https://github.com/idrori/mlfinalsQ. The code provided allows Note that grading machine outputs were done manually, so none of the provided code will produce those.

B Generating New Questions

Table 2: Generating new questions: example of a new question for each topic automatically generated and the closest question in the dataset based on cosine similarity of the questions embeddings.

Topic	Question	Similarity
Regression	Generated Question: “We’re given a data set $D = \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n}$, where $x^{(i)} \in R^d$ and $y^{(i)} \in R$. Let X be a $d \times n$ matrix in which the $x^{(i)}$ are the columns and let Y be a $1 \times n$ vector containing the values of $y^{(i)}$. Using the ordinary least-squares formula, we can compute $W_{ols} = \left(XX^T\right)^{-1} XY^T$ Using ridge regression, we can compute $W_{ridge} = \left(XX^T + \lambda I\right)^{-1} XY^T$ We decide to try to use these methods to initialize a single-unit "neural network" with a linear activation function. Assume that XX^T is neither singular nor equal to the identity matrix, and that neither W_{ols} nor W_{ridge} is equal to $(0, 0, \ldots, 0)$. Consider a neuron initialized with W_{ridge}. Provide an objective function $J(W)$ that depends on the data, such that batch gradient descent to minimize J will have no effect on the weights, or argue that one does not exist." Closest Question: “We’re given a data set $D = \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n}$, where $x^{(i)} \in R^d$ and $y^{(i)} \in R$. Let X be a $d \times n$ matrix in which the $x^{(i)}$ are the columns and let Y be a $1 \times n$ vector containing the values of $y^{(i)}$. Using the analytical regression (ordinary least-squares) formula, we can compute $W_{ols} = \left(XX^T\right)^{-1} XY^T$ Using ridge regression, we can compute $W_{ridge} = \left(XX^T + \lambda I\right)^{-1} XY^T$ We decide to try to use these methods to initialize a single-unit "neural network" with a linear activation function and no offset: $h(x; W) = W^T x.$ Assume that XX^T is invertible and not equal to the identity matrix, and that neither W_{ols} nor W_{ridge} is equal to $(0, 0, \ldots, 0)$. Note also that we are not using an explicit offset/bias term. Rory has solved many problems from this particular domain before and the solution has typically been close to $W^* = (1, \ldots, 1)^T$. Define an objective function $J(W)$ that we could minimize in order to obtain good estimates for Rory’s next problem, even with very little data."	0.945

Table 2: Generating new questions: example of a new question for each topic automatically generated and the closest question in the dataset based on cosine similarity of the questions embeddings.

Topic	Question	Similarity
Classifiers	Generated Question: “Consider a binary classification problem with two classes, +1 and -1. Assume that the training data is linearly separable. Consider the following two models: Model 1: $g(x) = sgn(w^T x)$ Model 2: $g(x) = sgn(w^T x + b)$ Assume that the training data is linearly separable. Which model is more likely to overfit?" Closest Question: “In some cases, we will have a validation set in addition to training and test sets. Assume the validation set is approximately the same size as the test set. This validation set is often used to tune hyperparameters such as λ. Imagine we have trained a classifier using regularization, with λ chosen based on performance on the training set. Which will have the highest accuracy the training set, the validation set or the test set?"	0.782
Logistic Regression	Generated Question: “In logistic regression, we use the sigmoid function to "squish" the output of our linear function to be between 0 and 1. What is the output of the sigmoid function when the input is 0?" Closest Question: “Consider the simplest of all neural networks, consisting of a single unit with a sigmoid activation function: $h(x; w = \sigma(w_0 + w_1 x)$ where $\sigma(z) = (1 + exp(-z))^{-1}$ Let’s start with a classifier defined by $w_0 = -1$ and $w_1 = 1$. Which range of input values x are classified as positive? Which as negative?"	0.810
Features	Generated Question: “You are working on a project to predict how successful a given Kickstarter campaign will be, based on information about the campaign (such as the amount of money pledged, the number of backers, the category of the project, the country where the project is based, and so forth). One of the features you want to use is the project’s title. What is a good way to encode the project’s title for use in a machine-learning predictor?" Closest Question: “Mac O’Larnin is considering selling an app on Frugal Play. You have a friend with inside info at Frugal, and they’re able to share data on how previous apps have performed on the store. Mac decides that he will learn a neural network with no hidden layer (i.e., consisting only of the output layer). He needs help in figuring out the precise formulation for machine learning. What is the best way to encode the app characteristic ’Suitable for people ages (2–4, 5–10, 11–15, 16 and over)’ as a feature for an input to the neural network? Choose from among the following: multiple unary features (one-hot encoding), multiple binary features (thermometer encoding), an integer or real-valued feature. Also give the exact function that maps each input to its corresponding feature(s)."	0.761

Table 2: Generating new questions: example of a new question for each topic automatically generated and the closest question in the dataset based on cosine similarity of the questions embeddings.

Topic	Question	Similarity
Loss Functions	Generated Question: “We have a data set of images, each of which is either a picture of a dog or a picture of a cat. We want to train a binary classifier to distinguish between the two classes. Which of the following is a better choice of loss function for this task: (a) The 0-1 loss, which gives a loss of 1 if the classifier predicts the wrong class and a loss of 0 otherwise. (b) The cross-entropy loss, which is given by -log(p) if the correct class is predicted and -log(1-p) otherwise, where p is the predicted probability of the correct class. (c) The hinge loss, which is given by max(0, 1-s) if the correct class is predicted and max(0, 1+s) otherwise, where s is the predicted score for the correct class." Closest Question: “Consider the following 2D dataset in (x,y) format: ((1,-1), +1), ((1,1), +1), ((1,2.5),+1), ((2,-2),-1), ((2,1),+1),((2,3),+1),((5,-1),-1),((5,-2),-1). Break ties in distance by choosing the point with smaller x_1 coordinate, and if still tied, by smaller x_2 coordinate. Compute the leave-one-out cross validation accuracy of the 3-nearest-neighbor learning algorithm on this dataset."	0.790
Neural Networks	Generated Question: “Consider a neural network with two hidden layers, each with two units. The input layer has two units, and the output layer has one unit. The hidden layers use the sigmoid activation function, and the output layer uses the linear activation function. The weights from the input layer to the first hidden layer are w11 = 1, w12 = 1, w21 = 1, and w22 = 1. The weights from the first hidden layer to the second hidden layer are w11 = 1, w12 = 1, w21 = 1, and w22 = 1. The weights from the second hidden layer to the output layer are w11 = 1, w21 = 1. The bias terms are all zero. What is the output of the neural network for the input x1 = 1, x2 = 1?" Closest Question: “A neural network is given as $Z^1 = X * W^1$, $A^1 = f1(Z^1)$, $Z^2 = W^2 * A^1$, $\hat{y} = f^2(Z^2)$. Specifically, the input X is a 4×1 column vector, $\hat{y}$ is a 1×1 scalar. W^2 is a 3×1 vector. We also know that, $Z^1 = (W^1)^T X$ and $Z^2 = (W^2)^T A^1$. What are the dimensions of Z^2?"	0.880
CNNs	Generated Question: “Suppose we have a 3x3 image and we use a 2x2 filter with stride 1. What are the dimensions of the output image?" Closest Question: “A neural network is given as $Z^1 = X * W^1$, $A^1 = f^1(Z^1)$, $Z^2 = W^2 * A^1$, $\hat{y} = f^2(Z^2)$. There is only one data point which is: $X = [1, 1, 1, 1]^T$ and $y = [1]$. If W^1 and W^2 are both matrices/vectors of all ones, what is the resulting Loss where the Loss = $(y - \hat{y})^2$?"	0.895

Table 2: Generating new questions: example of a new question for each topic automatically generated and the closest question in the dataset based on cosine similarity of the questions embeddings.

Topic	Question	Similarity
MDPs	Generated Question: “Consider an MDP with four states, called A, B, C, and D, and with two actions called Move and Stay. The discount factor $\gamma = 0.9$. Here is a reminder of the Q-learning update formula, based on experience tuple (s, a, r, s') : $$Q(s,a) := (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q\left(s', a'\right)\right)$$ Let $\alpha = 1$. Assume we see the following state-action-reward sequence: A, Move, 0 B, Move, 0 C, Move, 1 A, Move, 0 B, Move, 0 C, Move, 1 With Q-values all starting at 0, we run the Q-learning algorithm on that state-action sequence. Provide the q-learning value for Q(C, Move)." Closest Question: “Consider an MDP with four states, called A, B, C, and D, and with two actions called Move and Stay. The discount factor $\gamma = 0.9$. Here is a reminder of the Q-learning update formula, based on experience tuple (s, a, r, s') : $$Q(s,a) := (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q\left(s', a'\right)\right)$$ Let $\alpha = 1$. Assume we see the following state-action-reward sequence: A, Move, 0 B, Move, 0 C, Move, 1 A, Move, 0 With Q-values all starting at 0, we run the Q-learning algorithm on that state-action sequence. Provide the q-learning value for Q(A, move)."	0.988
RNNs	Generated Question: “Consider the following RNN: $$s_t = tanh(w_1 x_t + w_2 s_{t-1} + b) \ \ ,$$ $$y_t = w_3 s_t + b_2 \ \ .$$ Assume $s_0 = 0$ and $b_2 = 0$. What values of w_1, w_2, w_3 and b would generate output sequence $$[0, 0, 0, 1, 1, 1, 1]$$ given input sequence $$[0, 0, 0, 1, 0, 1, 0]$$ " Closest Question: “Ronnie makes a simple RNN with state dimension 1 and a *step* function for f_1, so that $$s_t = step(w_1 x_t + w_2 s_{t-1} + b)$$ where $step(z) = 1$ if $z > 0.0$ and equals 0 otherwise, and where the output $$y_t = s_t \ \ .$$ Assuming $s_0 = 1$, we want to generate output sequence $[0, 0, 0, 1, 1, 1, 1]$ given input sequence $[0, 0, 0, 1, 0, 1, 0]$. Rennie thinks this is not possible using Ronnie’s architecture. Rennie makes an argument based on the relationships in the table above. Is Rennie right?"	0.907

Table 2: Generating new questions: example of a new question for each topic automatically generated and the closest question in the dataset based on cosine similarity of the questions embeddings.

Topic	Question	Similarity
Reinforcement Learning	Generated Question: “What is the tabular Q-learning update equation, based on experience tuple (s, a, r, s')?" Closest Question: “b is the bias for classifier network. What are dimensions of b for multi k-class classification?"	0.796
Clustering	Generated Question: “Suppose that we have a dataset with n data points, k clusters, and d features. After running the k-means algorithm, the within-cluster sum of squared errors (WCSS) is given by: $\frac{1}{n}\sum_{i=1}^{n}\lVert\mathbf{x}_i - \mu_{y_i}\rVert^2$ where y_i is the cluster label of the ith data point, and μ_{y_i} is the cluster center associated with the ith data point. The within-cluster sum of squared errors (WCSS) is a measure of how well the clusters fit the data. Suppose that we have two datasets, X_1 and X_2, where X_1 has n_1 data points and X_2 has n_2 data points. We run the k-means algorithm on both datasets. We find that the WCSS for X_1 is smaller than the WCSS for X_2. Does this imply that the clusters for X_1 are better than the clusters for X_2? Why or why not?" Closest Question: “Consider the following 2D dataset in (x,y) format: ((1,-1), +1), ((1,1), +1), ((1,2.5),+1), ((2,-2),-1), ((2,1),+1),((2,3),+1),((5,-1),-1),((5,-2),-1). We will construct a tree using a greedy algorithm that recursively minimizes weighted average entropy. Recall that the weighted average entropy of a split into subsets A and B is: (fraction of points in $A) \cdot H\left(R_{j,s}^{A}\right) + ($ fraction of points in $B) \cdot H\left(R_{j,s}^{B}\right)$ where the entropy $H\left(R_m\right)$ of data in a region R_m is given by $H\left(R_m\right) = -\sum_k \hat{P}_{mk} \log_2 \hat{P}_{mk}$. The $\hat{P}_{mk}$ is the empirical probability, which is in this case the fraction of items in region m that are of class k. Some facts that might be useful to you: H(0) = 0, H(3/5) = 0.97, H(3/8) = 0.95, H(3/4) = 0.81, H(5/6) = 0.65, H(1) = 0. Draw the decision tree that would be constructed by our tree algorithm for this dataset. Clearly label the test in each node, which case (yes or no) each branch corresponds to, and the prediction that will be made at each leaf. Assume there is no pruning and that the algorithm runs until each leaf has only members of a single class."	0.767

Table 2: Generating new questions: example of a new question for each topic automatically generated and the closest question in the dataset based on cosine similarity of the questions embeddings.

Topic	Question	Similarity
Decision Trees	Generated Question: “The Gini score is a measure of how often a randomly chosen element would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. It ranges from 0 to 1, with 0 meaning that there are no mislabeled elements and 1 meaning that the subset is perfectly mixed. Explain whether or not it would be a good idea to use the Gini score as a scoring function for pruning decision trees." Closest Question: “There are different strategies for pruning decision trees. We assume that we grow a decision tree until there is one or a small number of elements in each leaf. Then, we prune by deleting individual leaves of the tree until the score of the tree starts to get worse. The question is how to score each possible pruning of the tree. Here is a definition of the score: The score is the percentage correct of the tree on a separate validation set. Explain whether or not it would be a good idea and give a reason why or why not."	0.867