140.644: Statistical Machine Learning **Name:** _____

Instructor: Yiqun Chen TA: Junwei (Ivy) Sun

Quiz 6 — Practice — Deep Learning and Non-Tabular Data

Not graded. Bring questions to class.

This is a shortened sample so you know what to expect. The real quiz has **4 multiple choice** and **2 questions in Part B**, and is worth 100 points; this one has 2 and 2, and the matching questions use fewer runs and methods. The style and the difficulty are the same. No calculator is needed.

Part A: Multiple Choice [20 points]

Circle the single best answer. **(10 points)** each.

1. We minimize $L(\theta) = (\theta - 2)^2$, whose derivative is $L'(\theta) = 2(\theta - 2)$, by gradient descent, $\theta \leftarrow \theta - \eta\, L'(\theta)$, with learning rate $\eta = 0.25$. We start at $\theta = 0$. What are the next two values of θ?

   (A) 2, then 2: one step reaches the bottom and stays

   (B) 1, then 1.5: each step goes half of the remaining way

   (C) 1, then 2: two equal steps reach the bottom

   (D) -1, then -2.5: the steps run away from the bottom

2. We compare the review "the food was delicious" with two others by cosine similarity: (i) "the meal was tasty" and (ii) "the food was disgusting". Which review is closer to it when each review is a row of word counts, and which when each review is a sentence embedding from a pretrained language model?

   (A) (i) under both

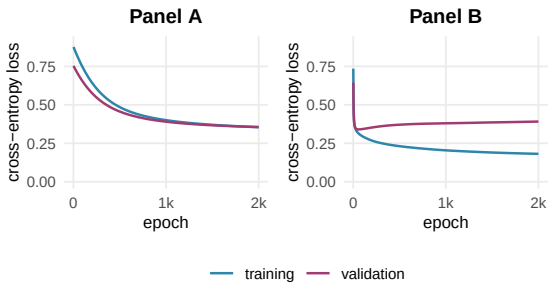
   (B) (ii) under word counts, and (i) under the sentence embedding

   (C) (i) under word counts, and (ii) under the sentence embedding

   (D) (ii) under both

Part B: Reading Training Runs and Leaderboards [40 points]

1. **(20 points)** The same network, with one hidden layer of 64 units, was trained twice on the same 150 training observations, for 2,000 epochs each. After every epoch, the loss was recorded on the training set and on a separate validation set. The two runs differ only in the learning rate, which was 0.0005 or 0.05.



Match each panel to the learning rate most likely to have produced it, and say what you read off the plot to decide.

Panel A	Panel B

2. **(20 points)** The data are 18,846 posts from 20 online discussion boards. Each post is one observation, and the text of the post is the only input. The outcome is binary: whether the post was made on one of the four recreation boards (cars, motorcycles, baseball, hockey), which is true for about 21% of posts. The 11,314 earlier posts are the training set, and the 7,532 later posts are the held-out set.

   Two classification methods were each trained on 100, 1,000, and all 11,314 training posts, and every fit was scored on the same held-out posts. The table shows the held-out AUC.

Held-out AUC	**100 posts**	**1,000 posts**	**11,314 posts**	**Method number**
Method A	0.952	0.973	0.980	
Method B	0.724	0.926	0.971	

   Each row is exactly one of the following methods. Match each row to the method most likely to have produced it, write the method's number in the last column, and in one or two sentences say which numbers in the table decided it.

   (1) **TF-IDF, then logistic regression.** Each post becomes a row of word counts, with one column per word. A word that appears in most posts is given less weight. Logistic regression is then fitted on those columns.

   (2) **Pretrained language model, then logistic regression.** Each post is passed through a language model that was trained beforehand on a large collection of general text. The language model is not updated. It returns a vector of 384 numbers for each post, and logistic regression is fitted on those 384 columns.