Practice Mid-term exam, CS570 Deep Learning Spring 2022

Name: ______________________________ StudentID: ______________________________ 11 Mar

1 Learning a linear model

Poisson regression is a machine learning problem where the output/label $y \in \{0, 1, 2, \dots\}$ is a non-negative integer, and the input/features $\mathbf{x} \in \mathbb{R}^p$ is a real vector. The loss function we use in this case is $\ell(\hat y, y) = \exp(\hat y) - y\hat y$ where $\hat y \in \mathbb{R}$ is a predicted score. In a linear model we use predicted scores defined by $\hat y = f(\mathbf{x}) = \beta + w^T\mathbf{x}$ where the linear model parameters are $\mathbf{w} \in \mathbb{R}^p$, a vector of p weights, and $\beta \in \mathbb{R}$, an intercept. Give details of a gradient descent algorithm with early stopping regularization which we could use to learn the parameters of the linear model.

1. how to split train data.

2. loss function to minimize via gradient descent (and what set it is defined on).

3. how to compute gradient and parameter updates, including expressions for gradient of loss with respect to weights and intercept.

4. when to stop the gradient descent algorithm.

2 Asymptotic complexity

Assume you have a train data set with n rows/observations and p columns/features that you use to make predictions using both a linear model and nearest neighbors. What is the asymptotic time/space complexity to run fit/predict methods for each learning algorithm? Write big O notation of fit/predict methods in terms of n and p (ignore time/space it takes for hyper-parameter selection via cross-validation)

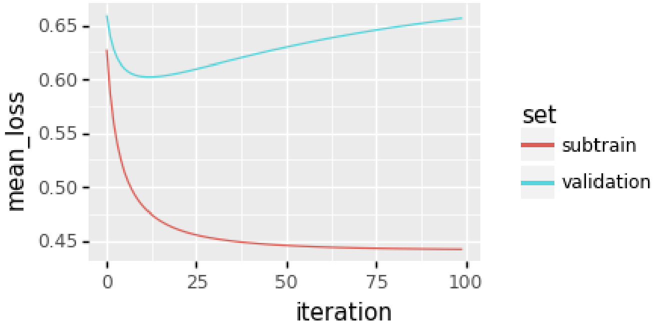
Learner	method	what it does	time	space
Linear model	fit			
	predict			
Nearest Neighbors	predict			
	fit			

3 Interpreting results

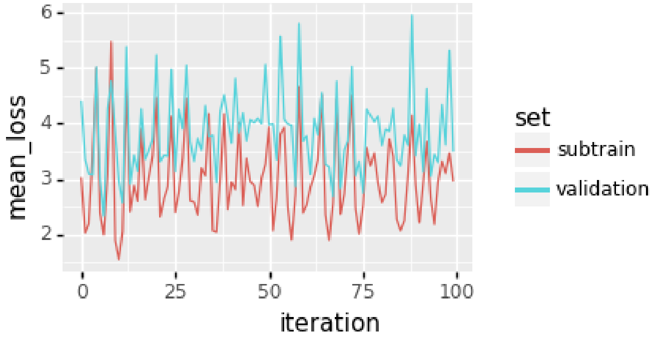
In each of the three scenarios below a linear model was learned using gradient descent with a constant step size.

1. For each of the three plots, indicate if the step size was too big, too small, or if it is a reasonable size, and explain why.

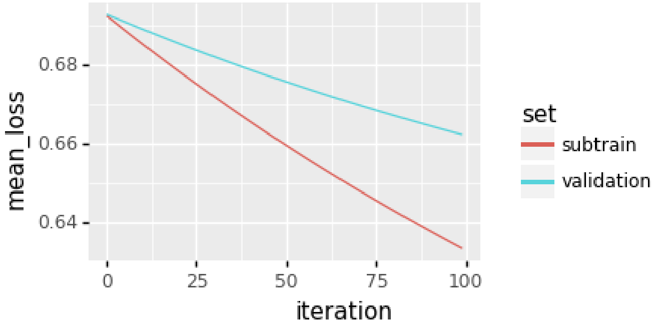
2. Draw a vertical line on one of the three plots at the number of iterations you should select to get best prediction accuracy, and briefly explain why.



A)



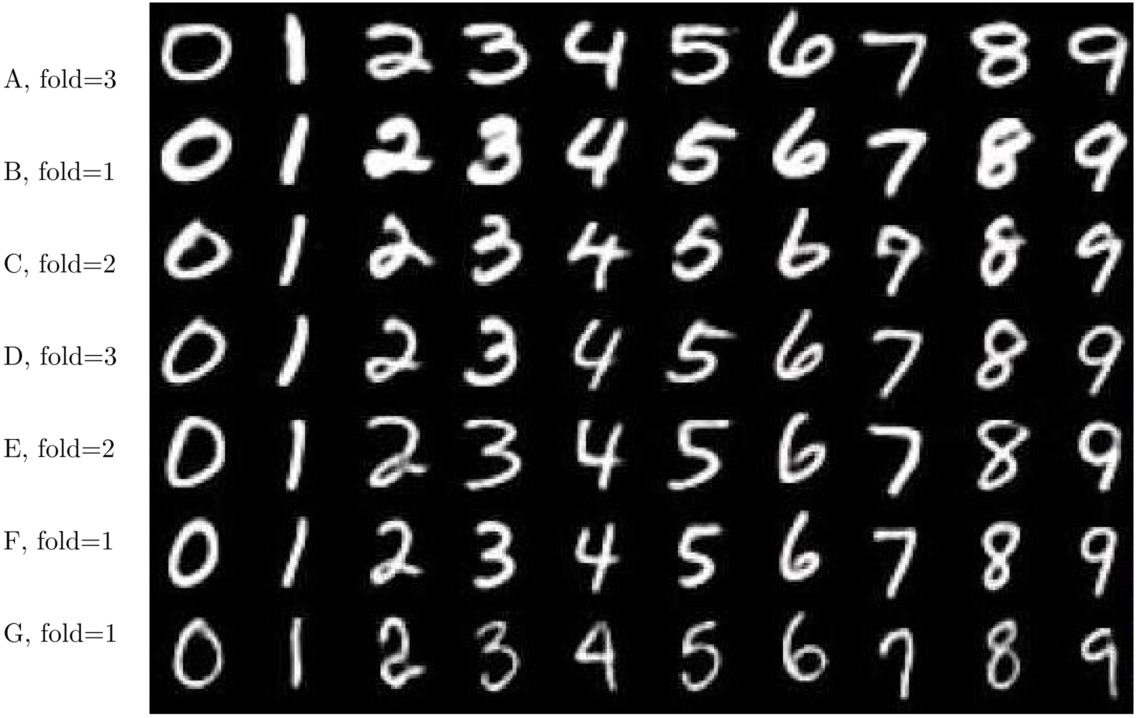
B)



C)

4 K-fold cross-validation

The image below represents a data set with 70 observations, one for each individual image of a digit. Say we want to determine which of several different machine learning models (e.g. linear model with early stopping, nearest neighbors, etc) is most accurate in these data. To do that we perform 3-fold cross-validation. Fold ID numbers $\in \{1, 2, 3\}$ have been assigned to all observations/images in the corresponding row/letter.



1. For fold/split 1 which observations/letters are the train set which are passed to the learning algorithm/function? ____________________

   For fold/split 1 which observations/letters are used for test set? (learning algorithm can not access these data, but the learned model is used to predict for them) ____________________

2. For fold/split 2. Train set = ____________________, Test set = ____________________.

3. For fold/split 3. Train set = ____________________, Test set = ____________________.

4. Now assume that we are in the context of fold/split 1. Your learning algorithm has access to all of the data in the train set. Your learning algorithm uses 2-fold cross-validation internally to select the best regularization hyper-parameter (steps, neighbors, penalty, etc). To do these subtrain/validation splits, randomly assign each observation/letter in the train set to a new/internal fold ID.

   Fold 1 = ____________________, Fold 2 = ____________________

5. Now assume that your function has computed MeanValidationLoss(η), the mean validation loss over both validation folds, for r regularization parameters $\eta_1, \dots, \eta_r$. What is the best model/regularization parameter η^* that you should select to make the final predictions on the test set for fold 1?

   $\eta^* =$ ____