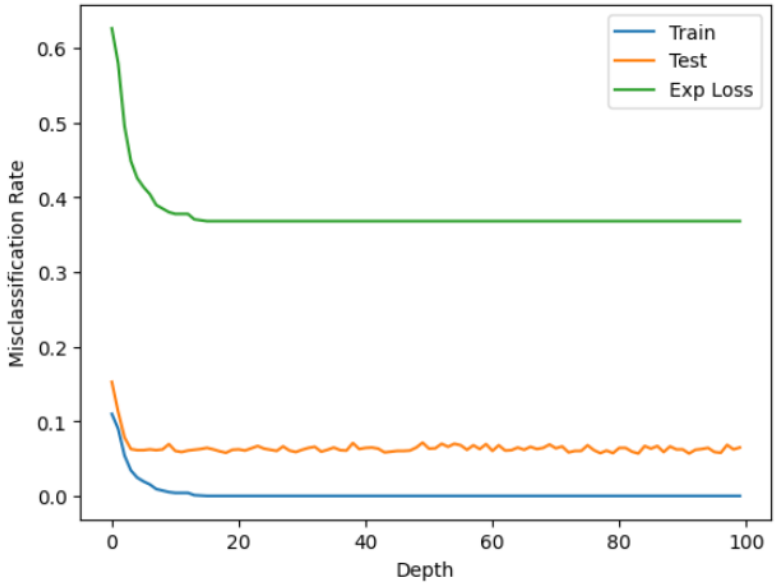
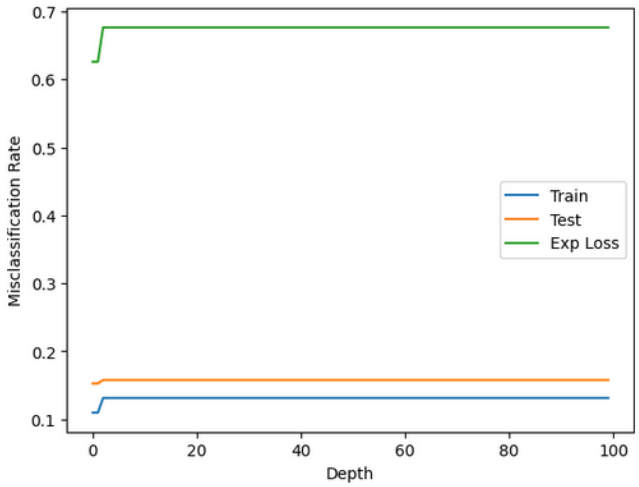
1. Machine learning fundamentals
    a. Decision trees can be used to return a new data label by further dividing classified data samples. The most important features/samples can be returned by measuring which classifying feature generated the most information gain. For example, if a decision tree was used to classify farm animals, at the leaf node that determines whether or not an animal is a chicken, the model can further analyze differing features based on a chicken's gender. This would provide a new label - gender of the animal - in addition to the previously returned label - species of animal. Also in this example, information gain be directly calculated to return which features (fur, hoofs, sound, size) or samples (animals in the barn, animals in the coop, animals out in the field) are the most important.
    b. Linear regression can be used to return a new data label because the learned model has been adjusted to minimize loss based on the training data.The most important features can be returned by calculating which features have the most correlation. For example, if a linear regression model was used to calculate house prices based on the many features of each house sample, new data labels could be generated based on how the regression was adjusted for current times. The learned model would be trained on past data, and adjusted to predict future house prices based on supplement data such as inflation and interest rates. This would create data labels for specific adjustments, like housing prices in 2002 vs 2008. The most important features depend on the largest relationship between the two features. Using correlations statistics, each feature can be given a linear correlation score from -1 to 0 - where the greater the score, the larger the relationship with 0 representing no relationship. For each feature, calculate its correlation score with every single feature. Then check how that feature fits within the data to result in a label. The features with the highest correlation score and with the greatest impact on the data to result in a label are the most important features.
    c. Graphical Models can be used to return a new data label because a learned graphical model uses probabilities to predict possible outcomes. The most important features can be returned by calculating which have the greatest effect when calculating probabilities of following events. For example, if a graphical model is used to predict whether or not it will rain for the next several days, the probability of it raining the next day can be used to generate a new label of whether or not it will be sunny, as opposed to raining. The probabilities of it raining are already utilized, and so to generate this new label would simply require extending those probabilities to calculate an opposite event. An event like whether or not it rained the previous day may have a greater effect on the probability of it raining than another event such as wind strength. In these cases,

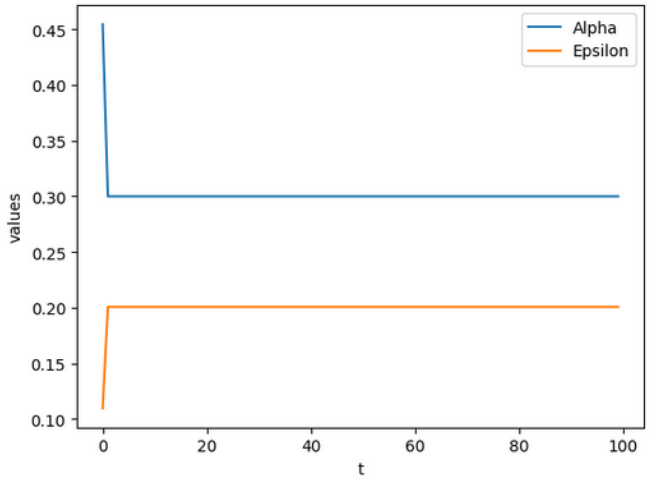
each calculation of differing data samples can be examined to see how differing probabilities of each event has the greatest effect on the outcome.

2. Adaboost
    a. Deep trees:



        i. smallest train misclassification rate: 0.0
        ii. smallest test misclassification rate: 0.055248618784530384
        iii. smallest train exponential loss: 0.3678794411714424
    b. Boosted decision stumps

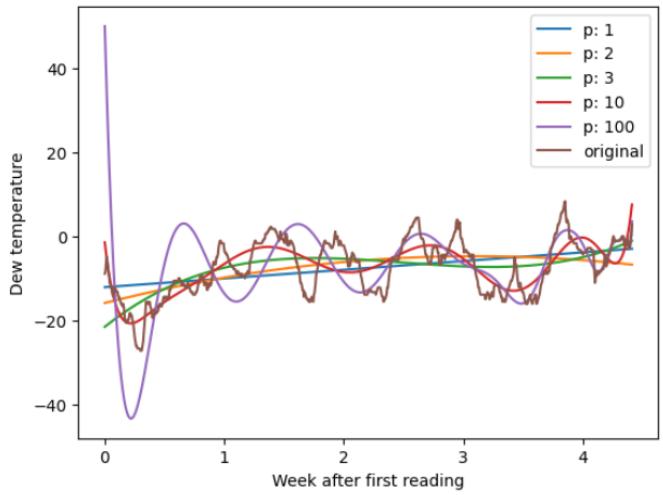




i. smallest train misclassification rate: 0.10980592441266598
ii. smallest test misclassification rate: 0.15268709191361124
iii. smallest train exponential loss: 0.6259675480492947
iv. Technically speaking, compared to the deep decision tree, the boosted decision stumps should perform better because adaboost utilizes a collection of stumps rather than a single boosted tree. There should be more accurate predictions at the cost of the likelihood of overfitting. Now, whether or not my flawed implementation can accurately reflect this is a different matter entirely.

c. Epsilon and Alpha
   i. Alpha is really large and positive when Epsilon is really small. Alpha is really large and negative when Epsilon is really big. Alpha is close to zero when Epsilon is close to 0. Boosting utilizes many weak classifiers to generate a strong classifier. In terms of weighted performance, boosting will improve when the number of classifiers increases.

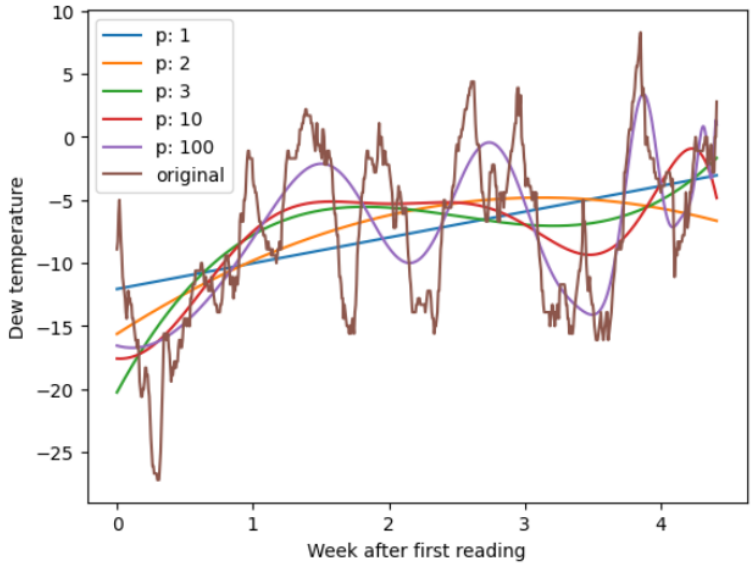
3. Polyfit via linear regression
    a. Linear Regression



i. The normal equation is: $x^T x\theta - x^T y = 0$
$\therefore x^T x\theta = x^T y$

ii. P = 1 is basically a straight line and does not fit the original data very well. P = 2 and P =3 are slightly better as they have curved behavior but it still is not a very good fit for the original data. P = 10 is a good fit for the original data of all orders of P. P = 100 is the best fit for the original data, although in my generated plot, the beginning appears erratic. As the number of weeks increases, the line for P = 100 closely follows that of the original data.

    b. Ridge Regression



i. The normal equation is: $x^T x\theta + \rho I\theta - x^T y = 0$

$\therefore x^T x\theta + \rho I\theta = x^T y$

ii. P =1 is again a straight line and therefore not a very good fit. P = 2 is a single curve and P = 3 is made up of two curves, making these better but still mediocre fits for the original data. P = 10 is closer to the original data but P = 100 again matches the original data well and is the best fit.

c. https://piazza.com/class/l3qfxflla3b1jx/post/129