1 Written Exercises: Analyzing Naive Bayes

Problem 1
Q1 Prove Bernoulli Naive Bayes has a linear decision boundary

Solution
Steps:

1. $\frac{P(y=1|x_1,...,x_d)}{P(y=0|x_1,...,x_d)} = \frac{\theta_1 \prod_{i=1}^{d} \theta_{i1}^{x_i}(1-\theta_{i1})^{1-x_i}}{\theta_0 \prod_{i=1}^{d} \theta_{i0}^{x_i}(1-\theta_{i0})^{1-x_i}}$
2. $\frac{\theta_1 \prod_{i=1}^{d} \theta_{i1}^{x_i}(1-\theta_{i1})^{1-x_i}}{\theta_0 \prod_{i=1}^{d} \theta_{i0}^{x_i}(1-\theta_{i0})^{1-x_i}} > 1$
3. $log(\frac{\theta_1 \prod_{i=1}^{d} \theta_{i1}^{x_i}(1-\theta_{i1})^{1-x_i}}{\theta_0 \prod_{i=1}^{d} \theta_{i0}^{x_i}(1-\theta_{i0})^{1-x_i}}) > log(1)$
4. $log(\theta_1 \prod_{i=1}^{d} \theta_{i1}^{x_i}(1-\theta_{i1})^{1-x_i}) - log(\theta_0 \prod_{i=1}^{d} \theta_{i0}^{x_i}(1-\theta_{i0})^{1-x_i}) > 0$
5. $log(\theta_1) + \sum_{i=1}^{d} log(\theta_{i1}^{x_i}(1-\theta_{i1})^{1-x_i}) - [log(\theta_0) + \sum_{i=1}^{d} log(\theta_{i0}^{x_i}(1-\theta_{i0})^{1-x_i})] > 0$
6. $log(\theta_1) + \sum_{i=1}^{d} log(\theta_{i1}^{x_i}\frac{(1-\theta_{i1})}{(1-\theta_{i1})^{1-x_i}}) - [log(\theta_0) + \sum_{i=1}^{d} log(\theta_{i0}^{x_i}\frac{(1-\theta_{i0})}{(1-\theta_{i0})^{1-x_i}})] > 0$
7. $log(\theta_1) + \sum_{i=1}^{d} log(\theta_{i1}^{x_i}) + log((1-\theta_{i1})) - log((1-\theta_{i1})^{x_i})) - [log(\theta_0) + \sum_{i=1}^{d} log(\theta_{i0}^{x_i}) + log((1-\theta_{i0})) - log((1-\theta_{i0})^{x_i}))] > 0$
8. $log(\theta_1) + \sum_{i=1}^{d} log(\theta_{i1}^{x_i}) + log((\theta_{i0})) - log((\theta_{i0})^{x_i})) - [log(\theta_0) + \sum_{i=1}^{d} log(\theta_{i0}^{x_i}) + log((\theta_{i1})) - log((\theta_{i1})^{x_i}))] > 0$
9. $log(\theta_1) - log(\theta_0) + dlog(\theta_{io} - dlog(\theta_{i1})[\sum_{i=1}^{d} log(\theta_{i1}^{x_i}) - log((\theta_{i0})^{x_i}))] - [\sum_{i=1}^{d} log(\theta_{i0}^{x_i}) - log((\theta_{i1})^{x_i}))] > 0$
10. $log(\theta_1) - log(\theta_0) + \sum_{i=1}^{d} log(\theta_{io}) - \sum_{i=1}^{d} log(\theta_{i1}) + [\sum_{i=1}^{d} log(\theta_{i1}^{x_i}) - log((\theta_{i0})^{x_i}))] - [\sum_{i=1}^{d} log(\theta_{i0}^{x_i}) - log((\theta_{i1})^{x_i})] > 0$
11. $log(\theta_1) - log(\theta_0) + \sum_{i=1}^{d} log(\theta_{io}) - \sum_{i=1}^{d} log(\theta_{i1}) + [\sum_{i=1}^{d} x_i log(\theta_{i1}) - x_i log(\theta_{i0})] - [x_i log(\theta_{i0}) - x_i log(\theta_{i1})] > 0$
12. $log(\theta_1) - log(\theta_0) + \sum_{i=1}^{d} log(\theta_{io}) - \sum_{i=1}^{d} log(\theta_{i1}) + [\sum_{i=1}^{d} x_i(log(\theta_{i1}) - log(\theta_{i0}) - log(\theta_{i0}) + log(\theta_{i1})] > 0$
13. $log(\theta_1) - log(\theta_0) + \sum_{i=1}^{d} log(\theta_{io}) - \sum_{i=1}^{d} log(\theta_{i1}) + [\sum_{i=1}^{d} x_i(2log(\theta_{i1}) - 2log(\theta_{i0}))] > 0$
14. ***Answer***:
    $log(\theta_1) - log(\theta_0) + \sum_{i=1}^{d} log(\theta_{io}) - \sum_{i=1}^{d} log(\theta_{i1}) + [\sum_{i=1}^{d} x_i(2log(\theta_{i1}) - 2log(\theta_{i0}))] > 0$
    $\rightarrow b + \sum_{i=1}^{d} w_i x_i > 0$

    Bias(b): $(log(\theta_1) - log(\theta_0) + \sum_{i=1}^{d} log(\theta_{io}) - \sum_{i=1}^{d} log(\theta_{i1}))$
    Weight(w_i): $(2log(\theta_{i1}) - 2log(\theta_{i0}))$

Problem 2
Q2 Duplicate Features in Naive Bayes

Solution
Prove: $P(y = 1 \mid X_1 = x_1, X_2 = x_2) > P(y = 1 \mid X_1 = x_1)$
Assumptions:

1. $P(y = 1 \mid X_1 = x_1) > P(Py = 0 \mid X_1 = x_1)$
2. $P(y = 1) = P(y = 0)$

3. $P(X_2 = x_2 \mid y) = P(X_1 = x_1 \mid y)$

Steps:

1. Full Formula for $P(y = 1 \mid X_1 = x_1) > P(Py = 0 \mid X_1 = x_1)$
2. $\frac{P(X_1=x_1|y=1)P(y=1)}{P(X_1=x_1|y=1)+P(X_1=x_1|y=0)} > \frac{P(X_1=x_1|y=0)P(y=0)}{P(X_1=x_1|y=1)+P(X_1=x_1|y=0)}$
3. $P(X_1 = x_1 \mid y = 1) > P(X_1 = x_1 \mid y = 0)$ [Simplify]
4. $1 > \frac{P(X_1=x_1|y=0)}{P(X_1=x_1|y=1)}$ [Moved left side to right]
5. Finding Full Formula for $P(y = 1 \mid X_1 = x_1, X_2 = x_2)$
6. $P(y = 1 \mid X_1 = x_1, X_2 = x_2) = \frac{P(X_1=x_1,X_2=x_2|y=1)P(y=1)}{P(X_1=x_1,X_2=x_2|y=1)P(y=1)+P(X_1=x_1,X_2=x_2|y=0)P(y=0)}$ [Bayes Rule]
7. $= \frac{P(X_1=x_1|y=1)^2P(y=1)}{P(X_1=x_1|y=1)^2P(y=1)+P(X_1=x_1|y=0)^2P(y=0)}$ [Naive Bayes Assumption and Assumption 3]
8. $= \frac{P(X_1=x_1|y=1)^2}{P(X_1=x_1|y=1)^2+P(X_1=x_1|y=0)^2}$ [Simplify]
9. Following the steps from 1-4, using the full formula for $P(y = 1 \mid X_1 = x_1, X_2 = x_2)$, we get $1 > (\frac{P(X_1=x_1|y=0)}{P(X_1=x_1|y=1)})^2$
10. Normalizing $P(y = 1 \mid X_1 = x1)$, we get $\frac{1}{1+(\frac{P(X_1=x_1|y=0)}{P(X_1=x_1|y=1)})}$
11. Normalizing $P(y = 1 \mid X_1 = x1, X_2 = x_2)$, we get $\frac{1}{1+(\frac{P(X_1=x_1|y=0)}{P(X_1=x_1|y=1)})^2}$
12. ***Answer***: Fractions with smaller denominators yield larger numbers. Since the denominator $1 + (\frac{P(X_1=x_1|y=0)}{P(X_1=x_1|y=1)})^2 < 1+(\frac{P(X_1=x_1|y=0)}{P(X_1=x_1|y=1)})$ for all numbers between (0, 1), the probability of $P(y = 1 \mid X_1 = x1, X_2 = x_2)$ is greater than $P(y = 1 \mid X_1 = x1)$

2 Implementing Back-propagation for Feed-forward Neural Network

Problem 3
Q3 Implementing the Backward Pass for a Linear Layer

Solution
Steps:

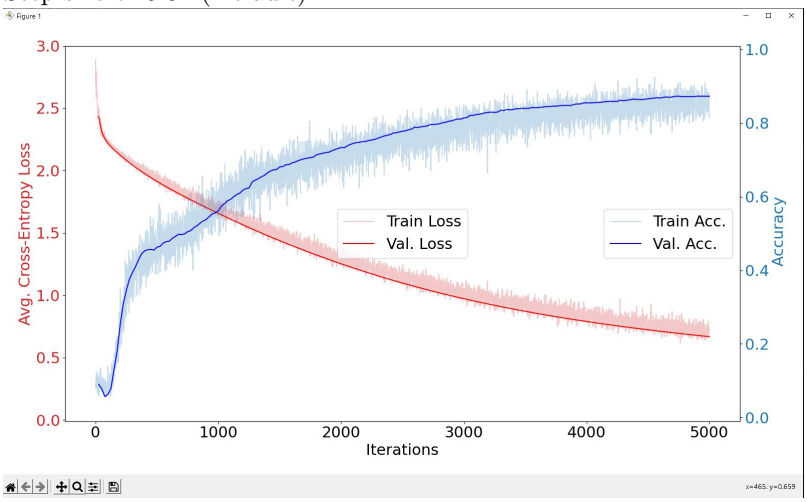
1. Completed

Problem 4
Q4 Learning Rate

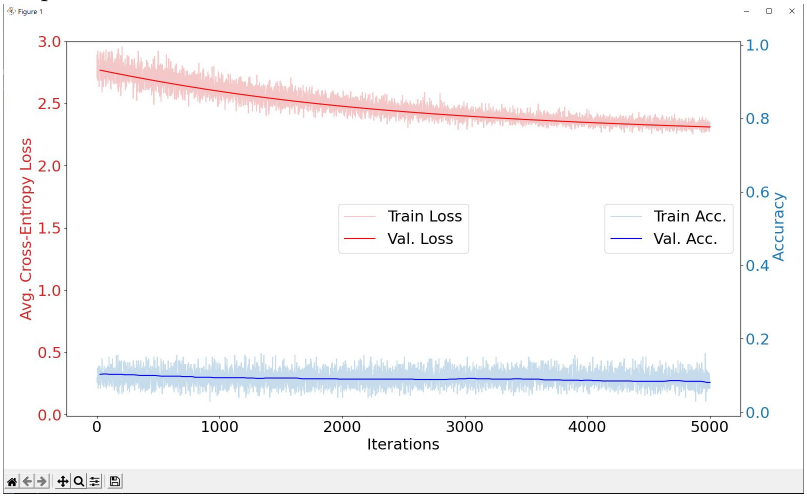
Solution

1. Plots

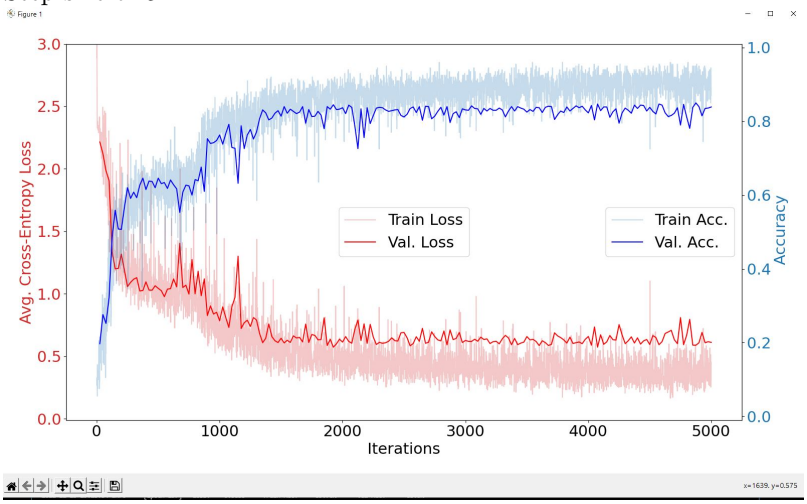
- Step size of 0.01 (Default)



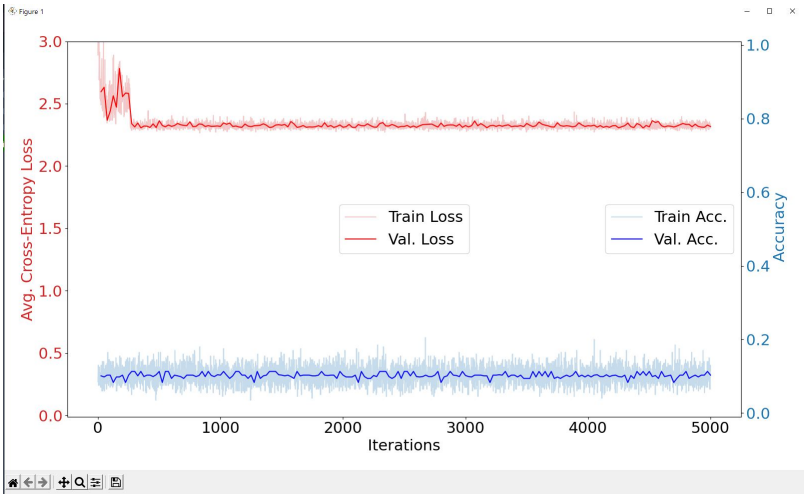
- Step size of 0.0001



- Step size of 5



- Step size of 10



2. (a) There are a couple differences I noticed when comparing the plots from various step sizes. The step sizes at each extreme end, 0.0001 and 10, are very inaccurate and have a high avg. cross entropy loss. Moreover, it appears that as the step sizes increase, the smoothness of the curves and their performance decreases.

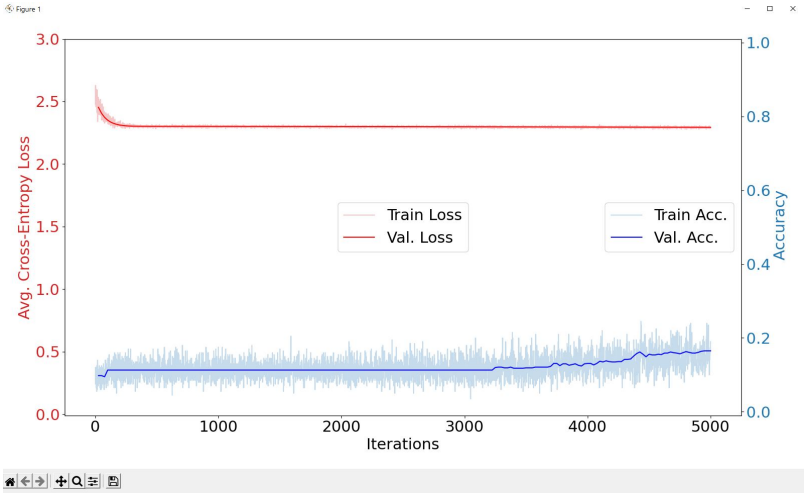
3. (b) If the max epochs were increased, I expect the avg. cross entropy loss and accuracy of the already performant curves to perform better. I also believe that increasing the max epochs would greatly improve curves with smaller step-sizes because they take longer to reach convergence.

Problem 5
Q5 ReLU's and Vanishing Gradients

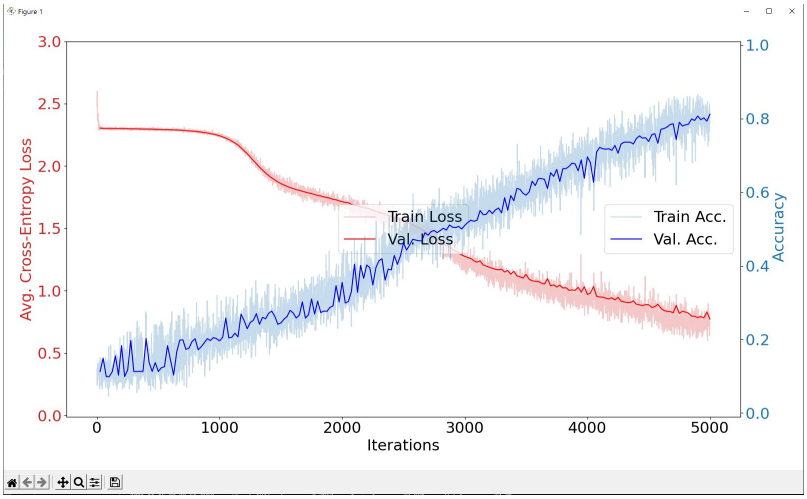
Solution

1. Plots

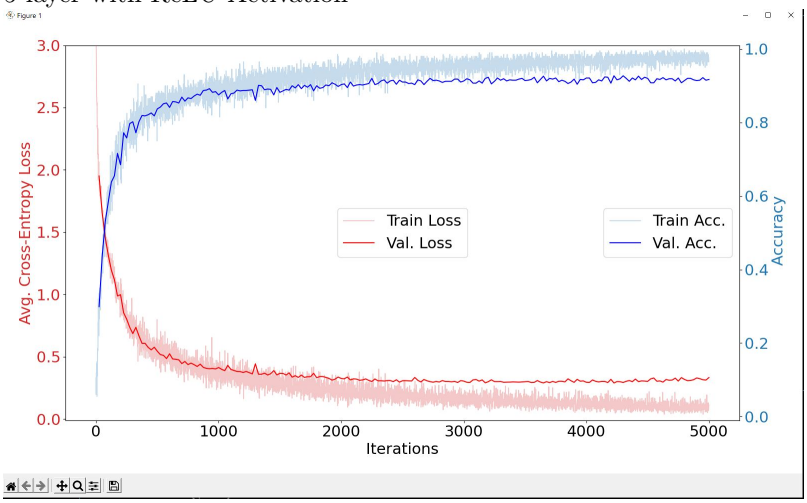
   - 5-layer with Sigmoid Activation (Default)



   - 5-layer with Sigmoid Activation with 0.1 step size



- 5-layer with ReLU Activation



2. (a) Between the 5-layer Sigmoid Activations with different step-sizes, it appears that the curve with the larger step-size was less smooth, formed a cross Loss and Accuracy data points, and had better performance than the curve with the smaller step-size, yielding a higher accuracy and lower loss at the end. Between the 5-layer Sigmoid Activation with 0.1 step size and the 5-layer ReLU Activation, the ReLU Activation curve was smoother, formed a cross much earlier, and was more performant than the former.

3. (b) The reason may be because the larger step-size increased the speed of the descent, allowing the curve to reach the local minimum faster than the smaller step-size over the same number of iterations.

4. (c) The ReLU Activation curve outperformed the Sigmoid Activation curve because, compared to the sigmoid's derivative containing exponents, its derivative is a constant. This greatly decreases the time it takes to compute the gradient. Moreover, the ReLU also converges faster than the sigmoid because the derivative of the sigmoid suffers from vanishing gradients. Since the derivative of the sigmoid causes the outputs to be smaller than the inputs, the gradient constantly gets smaller over each layer and approaches 0, causing sub-optimal training the further it propagates backwards.

Problem 6
Q6 Measuring Randomness

Solution

1. seed(102): 87.3%

   seed(1): 88.7%

- seed(10): 87.9%
- seed(100): 87.6%
- seed(1000): 89.1%
- seed(10000): 88.5%

2. It seems like the randomness makes the certainty of the conclusions in the previous questions weaker because it may be happen to use a certain seed that sets up the neural network to perform better or worse than other seeds. Therefore, it would be wise to average the performances across a large number of seeds to accurately measure performance.

Problem 7
Q7 Evaluating Cross Validation

Solution

1. For modifications, I changed the np.random.seed, decreased the batch size to 10, increased the max_epochs to 1000, increased the width of layers to 100, and set the activation to ReLU.

3 Debriefing

Problem 1
Approximately how many hours did you spend on this assignment?

Solution

1. About 8 hours.

Problem 2
Would you rate it as easy, moderate, or difficult?

Solution

1. Moderate-easy.

Problem 3
Did you work on it mostly alone or did you discuss the problems with others?

Solution

1. Worked on it alone.

Problem 4
How deeply do you feel you understand the material it covers (0%–100%)?

Solution

1. I felt I understood about 85 percent of the material covered.

Problem 4
Any other comments?

Solution

1. None.