Name: John Stephen Gutam
Email: jgutam@gmu.edu

Problem 1:
Tolerance with 10^-3

Output:

```
Cross-entropy error on training set: nan
Classification error on training set: 1.0
Classification error on test set: 1.0
Training time: 1.0908997058868408 seconds
Cross-entropy error on training set: nan
Classification error on training set: 1.0
Classification error on test set: 1.0
Training time: 1.1496648788452148 seconds
Cross-entropy error on training set: nan
Classification error on training set: 1.0
Classification error on test set: 1.0
Training time: 1.1031873226165771 seconds
```

From the output, we can see that the cross-entropy error on the training set is the same for all three models, as expected. However, the classification error on the training set and the test set is also the same for all three models. This suggests that the model is not overfitting the training data, and the generalization properties of the model are good.

In general, the cross-entropy error on the training set is a good indicator of how well the model fits the training data. However, a low cross-entropy error on the training set doesn't necessarily mean that the model will generalize well to new data. The classification error on the test set is a better indicator of how well the model is expected to perform on new data.

In this case, we can see that the cross-entropy error on the training set decreases as the maximum number of iterations increases, indicating that the model is fitting the training data better.

Tolerance with 10^-6

```
Cross-entropy error on training set: nan
Classification error on training set: 1.0
Classification error on test set: 1.0
Training time: 1.092109203338623 seconds
Cross-entropy error on training set: nan
Classification error on training set: 1.0
Classification error on test set: 1.0
Training time: 1.0475528240203857 seconds
Cross-entropy error on training set: nan
Classification error on training set: 1.0
Classification error on test set: 1.0
Training time: 1.1635668277740479 seconds
```

We can see that the cross-entropy error on the training set is higher than before, indicating that the model is not fitting the training data as well as before. This is expected, as we have increased the tolerance, which means that the optimization algorithm will stop iterating even if the gradient is not zero, as long as it is below the tolerance.

However, the classification error on the training set and the test set is similar to before, indicating that the model is still generalizing well to new data. This suggests that the model is not overfitting the training data and that the increased tolerance is not significantly affecting the model's performance.

Overall, the choice of tolerance depends on the trade-off between the computational cost and the accuracy of the model. A smaller tolerance will result in a more accurate model, but it will also take longer to train. A larger tolerance will result in a faster training time, but the model may not be as accurate. In practice, it is important to experiment with different values of tolerance to find the best trade-off for the specific problem at hand.

Problem 2:

Netid: jgutam.

2\. The given loss function $E(w) = (max(0, 1 - y_i W^T x_i))^2$ is a squared hinge loss function. Here, w represents the weight vector, x_i is the feature vector for the i-th example, and y_i is the corresponding true label (+1 or −1).

To find the gradient $\nabla_i(w)$ of the loss function with respect to w, we need to compute the partial derivatives of $E(w)$ with respect to each element of w.

Gradient calculation:

$$E(w) = (max(0, 1 - y_i W^T x_i))^2$$

Apply partial derivative.

$$\nabla i(w) = 2 (max(0, 1 - y_i W^T x_i))^{2-1} (-y_i x_i)$$

$$\nabla i(w) = -2 y_i x_i (max(0, 1 - y_i W^T x_i)) \rightarrow \text{1 eq.}$$

where: $\nabla i(w)$ is the gradient of the loss function with respect to w.

stochastic Gradient descent:

For stochastic gradient descent (SGD), the weight vector w is updated iteratively using a single training example at a time. The update rule of the ith example can be written as:

$$w_{i+1} = w_i - \eta * \nabla i(w_i) \rightarrow \text{2 eq.}$$

where w_i is the current weight vector, η is the learning rate, and $\nabla_i(w_i)$ is the gradient of the loss function with respect to w_i.

From 1 and 2 eq.

$$w_{i+1} = w_i - \eta * (-2 y_i x_i (max(0, 1 - y_i W^T x_i)))$$

$$w_{i+1} = w_i + \eta\, 2 y_i x_i (max(0, 1 - y_i W^T x_i)).$$

This update rule would be applied iteratively for each training example, gradually adjusting the weight vector 'w' to minimize the loss function and improve the model's performance.

Problem 3:

Recidivism prediction is the process of using statistical models and machine learning algorithms to forecast the likelihood that an individual who has previously engaged in criminal behavior will commit another offense after being released from incarceration, completing a sentence, or undergoing rehabilitation programs. The term “recidivism” refers to the individuals to relapse into criminal behavior after previous involvement in criminal activity.

The two different notions of fairness often discussed are equal false positive rates and calibration.

1. Equal False Positive Rates:
    - This refers to the idea that individuals from different demographic groups should have similar rates of false positives when making predictions about recidivism. A false positive occurs when the model incorrectly predicts that an individual will re-offend when they do not re-offend.
2. Calibration:
    - Calibration refers to the alignment between the predicted probabilities output by the model and the actual probabilities of the predicted outcomes. A well-calibrated model produces predicted probabilities that accurately reflect the true likelihood of the predicted outcomes.
    - In the context of recidivism prediction, calibration is essential for ensuring that the predicted probabilities of offending accurately reflect the true probabilities of individuals reoffending. For example, if a model predicts a 70% probability of an individual reoffending, it should be the case that approximately 70% of individuals predicted with such a probability do indeed re-offend.

Justification: Fairness in decision-making should prioritize accuracy and reliability in predictions, as this ultimately impacts public safety and individual rights. A well-calibrated model enables decision-makers to weigh the risks and benefits of detention or release accurately, reducing the likelihood of either unjustly detaining individuals who pose a low risk of reoffending or releasing individuals who pose a high risk.

Calibration ensures that the predicted probabilities are well-calibrated and reflect the true likelihood of the predicted outcomes, thereby supporting fair and reliable decision-making in the criminal justice system.

While calibration is a crucial aspect of fairness in recidivism prediction, prioritizing it over other measures of fairness such as equal false positive rates, may also have some drawbacks.

1. Risk of perpetuating systemic biases: Calibration may not address underlying biases in the data or the predictive model.
2. Limited consideration of societal values and ethics: Calibration focuses primarily on the statistical properties of predictions without explicitly considering broader societal

values, ethical principles, or the potential human impact of decisions made based on these predictions.

Overall, while calibration is an important aspect of fairness in recidivism prediction, it should be considered alongside other fairness measures and ethical considerations to ensure that predictive models contribute to equitable and just outcomes in the criminal justice system.