DSC 140A - Quiz 03

Name:
PID:

Quiz 03 Instructions. All technology is prohibited. Pen and paper only. You may not discuss with any classmates. Any prohibited use of notes or technology will result in a 0 and is a violation of Academic Integrity. You will have 50 minutes to complete your quiz. If you have any questions, please raise your hand and a TA will come assist you. Make sure to write your name and PID at the top of every page as clearly as possible. Explain/show your work.

Notation: Bold upper case letters $\mathbf{A}$ represent matrices.
Bold lower case letters $\mathbf{a}$ represent vectors.
Non-bolded lower case letters a represent scalars.
For a matrix $\mathbf{A}$, $\mathbf{a}_i$ indicates the ith row vector of matrix $\mathbf{A}$.

Name:
PID:

Problem 1.

Classifier Objective Functions: We learned of several possible loss functions and some of the problems or drawbacks of using these loss functions for a linear classifier. For each of the loss functions explain the drawback/problem of optimizing for this loss function.

a) Square Loss

Solution: Lecture 6 Slide 92 tells us that square loss for linear classification can be problematic because it penalizes points that are far away from the decision boundary even if they are correctly classified.

b) 0-1 Loss

Solution: Lecture 6 Slide 102 tells us that 0-1 loss is problematic because the gradient of the Risk can very easily be 0 everywhere, i.e. the Risk function is flat for long stretches, making gradient descent unusable.

c) Perceptron Loss

Solution: Lecture 7 Slide 4 tells us that Perceptron loss is problematic because the loss assigns no penalty to correctly classified points, no matter how close the point is to the decision boundary. This is bad because it can lead to many points near the decision boundary, i.e. overfitting.

Name:
PID:

Problem 2.

You have the following training data:

i	x_1	x_2	y_i
1	3	1	$+1$
2	-1	2	-1

Two candidate classifiers (with no intercept) are proposed:

- **Classifier A:** $\mathbf{w} = (1, 1)$
- **Classifier B:** $\mathbf{w} = (3, 0)$

a) For each classifier, check whether every training point satisfies the Hard-SVM constraint

$$y_i \left(\mathbf{w}^T \mathbf{x}^{(i)} \right) \geq 1$$

for all i. **Select all classifiers that meet the constraint for all points.**

☐ Classifier A

✓ Classifier B

b) For any classifier that satisfies all constraints, compute the margin width.

Solution: The margin width is

$$\frac{2}{\|\mathbf{w}\|}.$$

Compute the norm of $\mathbf{w}$:

$$\|\mathbf{w}\| = \sqrt{3^2 + 0^2} = 3.$$

Therefore, the margin width is

$$\frac{2}{\|\mathbf{w}\|} = \boxed{\frac{2}{3}}.$$

c) Which classifier better minimizes $\|\mathbf{w}\|$? Which classifier would the Hard-SVM select? **Select all that apply.**

✓ Classifier A better minimizes $\|\mathbf{w}\|$

☐ Classifier B better minimizes $\|\mathbf{w}\|$

☐ Hard-SVM would select Classifier A

✓ Hard-SVM would select Classifier B

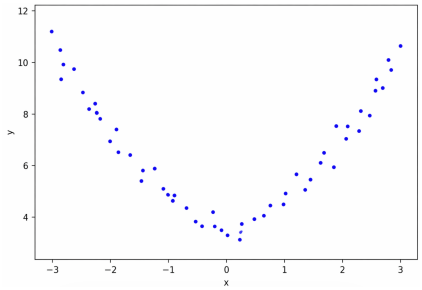
Name:
PID:

Problem 3.

A classmate proposes: *"We should always choose a very large C in a soft-margin SVM to make sure we misclassify as few training points as possible."* Which of the following statements are correct aout this reasoning? **Select all that apply.**

- ✓ A very large C forces the SVM to prioritize minimizing misclassifications, which can result in a narrower margin and a more sensitive decision boundary.
- ☐ A very large C is always the correct choice when the training data is linearly separable.
- ✓ If the training data contains noise or outliers, a very large C may cause the decision boundary to bend around them, hurting generalization to new data.
- ✓ A smaller C allows some training misclassifications but can produce a wider margin, which often generalizes better to unseen data.
- ☐ A very large C turns the soft-margin SVM into a hard-margin SVM, which always produces a better decision boundary regardless of the data.
- ☐ A very large C minimizes train misclassifications, which is what we care about the most.

Problem 4.



Given the above plot, Student A says: "Just train a linear predictor on the original feature." Student B says: "The feature map is essential — without it, no linear predictor can do well on this data." Which of the following statements are correct? **Select all that apply.**

- ☐ Student A is wrong because linear predictor cannot be trained on 2-dimensional data.
- ✓ Student B is more correct — the original data is not linearly separable, so a linear predictor will struggle without a feature map.
- ✓ A feature map can turn a non-linear pattern in input space into a linear pattern in feature space.
- ✓ After applying a feature map, we still train a linear model — just in the new feature space.
- ☐ Student A's approach could work perfectly if we simply collect more training data.
- ☐ Using a feature map means we cannot use empirical risk minimization.