DSC 140A - Midterm Version B

May 8, 2026

Name: SOLUTIONS

PID:

By signing below, you agree that you will behave honestly and fairly during and after this exam. You should not discuss any part of this exam with anyone who has not yet taken it.

Signature:

Name of student to your **left**:

Name of student to your **right**:

(Write "N/A" if a wall/aisle is to your left/right.)

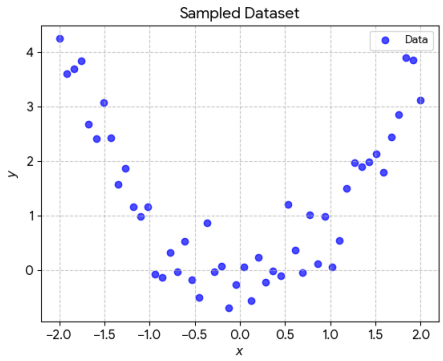
Instructions:

- Write your solutions to the following problems in the spaces provided.
- A double-sided A4 size cheatsheet is permitted.
- Write your name or PID at the top of each sheet in the space provided.
- Unfortunately, we cannot answer clarifying questions during the exam.
  - If you think a question is ambiguous (or has a bug), please write your assumptions beside the problem, and make a private post on Campuswire afterwards and we'll see if it needs to be addressed in grading.
- The exam is 9 pages in total (including this front page). Don't leave the last page empty.

(Please do not open your exam until instructed to do so.)

Problem 1. (2 points)

The following plot shows a y as a function of x. We want to fit a linear model to predict y by applying a feature map to x.



Which of the following feature maps $\phi(x)$ would be appropriate for performing linear regression to model the curvature of this data? *(Select all that apply)*

- ○ $\phi(x) = [1, \exp(-x^2/1000^2)]^T$
- ● $\phi(x) = [1, x, x^2]^T$
- ○ $\phi(x) = [1, \sin(100x)]^T$
- ○ $\phi(x) = [1, \exp(-(x-100)^2/0.5^2)]^T$
- ● $\phi(x) = [1, \exp(-x^2/1.5^2)]^T$

Solution: Explanation:

- $\phi(x) = [1, \exp(-x^2/1000^2)]^T$ is incorrect. A width of 1000 is so large that the function value is nearly 1.0 for all x in our range. This makes the feature almost indistinguishable from a constant bias term, failing to capture the quadratic trend.
- $\phi(x) = [1, x, x^2]^T$ is correct. This polynomial basis directly includes the quadratic feature needed to model the curvature.
- $\phi(x) = [1, \sin(100x)]^T$ is incorrect. A high-frequency sine wave oscillates far too rapidly to fit the smooth quadratic curve and would likely lead to a poor approximation.
- $\phi(x) = [1, \exp(-(x-100)^2/0.5^2)]^T$ is incorrect. This basis is centered at $x = 100$. For data points near $[-2, 2]$, the value of this feature will be effectively zero, so it provides no useful information for the model.
- $\phi(x) = [1, \exp(-x^2/1.5^2)]^T$ is correct. A Gaussian basis centered at the origin with a width comparable to the data range has significant curvature. Linear regression can assign it a negative weight to approximate the upward-opening quadratic bowl over this local interval.

Problem 2. (2 points)

Identify the True statements regarding L_1 and L_2 regularization (select all that apply):

- ○ The constraint region for L_2 regularization is shaped like a diamond, which encourages the solution to fall on the axes.
- ● Regularization techniques are designed to reduce overfitting by adding a penalty term that discourages overly complex models.

- ◯ Increasing the regularization parameter λ will always increase the training error, so we should keep the λ as small as possible.
- ● L_2 regularization can help solve issues with multicollinearity by ensuring the matrix $(\Phi^T\Phi + n\lambda I)$ is non-singular and well-conditioned.
- ◯ L_2 regularization is more effective at producing sparse weight vectors with many exact zeros than L_1 regularization.
- ◯ During Gradient Descent, L_2 regularization subtracts a fixed constant from every weight at each step, regardless of the weight's current magnitude.

Solution: Explanation of Choices:

- **Wrong:** The L_2 constraint is a circle/hypersphere. The L_1 constraint is the diamond-shaped region that encourages axis-aligned sparse solutions.
- **Correct:** Regularization penalizes overly complex models, often by discouraging large weights, which can reduce overfitting and improve generalization.
- **Wrong:** Larger λ often increases training error because it prioritizes smaller weights over perfectly fitting the training data, but the conclusion is wrong. We should choose λ by validation or cross validation, not simply make it as small as possible.
- **Correct:** Adding a positive multiple of the identity matrix to $\Phi^T\Phi$ makes the linear system better conditioned and can remove singularity issues caused by multicollinearity.
- **Wrong:** L_1 regularization, not L_2 regularization, is the one known for producing sparse weight vectors with many exact zeros.
- **Wrong:** L_2 regularization produces a shrinkage term proportional to the current weight value. Subtracting a fixed constant regardless of magnitude is more characteristic of the L_1 subgradient away from zero.

Problem 3. (2 points)

a) Let $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)}$ be n vectors in $\mathbb{R}^d$, and let K be the Gram matrix for this data set using the Gaussian RBF kernel. What type of object is K?

- ◯ A $d \times d$ matrix
- ◯ A scalar
- ◯ An n-dimensional vector
- ● An $n \times n$ matrix
- ◯ A d-dimensional vector

b) Let Φ be an $n \times k$ matrix, let λ be a scalar, let $\vec{y}$ be an n dimensional vector, and let I be the $k \times k$ identity matrix. What type of object is the following? You may assume that the matrix inverse exists.

$$(\Phi^T \Phi + n\lambda I)^{-1} \Phi^T \vec{y}$$

- ◯ An n-dimensional vector
- ◯ A $k \times k$ matrix
- ● A k-dimensional vector
- ◯ A scalar
- ◯ An $n \times n$ matrix

c) Let $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)}$ be n vectors in $\mathbb{R}^d$. Also let $\vec{w}$ be a vector in $\mathbb{R}^d$. What type of object is:

$$\sum_{i=1}^{n} \left(\vec{x}^{(i)} \cdot \vec{w} \right) \vec{x}^{(i)}?$$

- ◯ An $n \times n$ matrix
- ◯ A scalar
- ◯ A $d \times d$ matrix
- ● A vector in $\mathbb{R}^d$
- ◯ A vector in $\mathbb{R}^n$

d) Let $(\vec{x}^{(1)}, y_1), \ldots, (\vec{x}^{(n)}, y_n)$ be a labeled data set where $\vec{x}^{(i)} \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Let $\vec{\phi} : \mathbb{R}^d \to \mathbb{R}^k$ be a feature map, and suppose $\vec{w}$ is a vector in $\mathbb{R}^k$. Define:

$$R(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \vec{w} \cdot \vec{\phi}(\vec{x}^{(i)}))^2.$$

If $\vec{w}^*$ is a vector in $\mathbb{R}^k$, what type of object is the *gradient* of R evaluated at $\vec{w}^*$?
Caution! Note that the question is asking what type of object the *gradient* at $\vec{w}^*$ is, not what type of object $R(\vec{w}^*)$ is.

- ◯ A vector in $\mathbb{R}^n$
- ◯ A $d \times k$ matrix
- ◯ A scalar
- ● A vector in $\mathbb{R}^k$
- ◯ A vector in $\mathbb{R}^d$
- ◯ An $n \times n$ matrix

Problem 4. (2 points)

Given two d-dimensional vectors $\vec{u}$ and $\vec{v}$, we define the product kernel $\kappa(\vec{u}, \vec{v})$ as:

$$\kappa(\vec{u}, \vec{v}) = \sum_{i=1}^{d} (u_i \cdot v_i)$$

Suppose a prediction function $H(\vec{x})$ is trained using kernel ridge regression on the following 2D dataset:

$$\vec{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad y_1 = 1$$
$$\vec{x}_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \quad y_2 = -1$$
$$\vec{x}_3 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \quad y_3 = 1$$

Given that the solution to the dual problem (the dual weights) is:

$$\vec{\alpha} = \begin{bmatrix} 1.0 \\ -0.5 \\ 2.0 \end{bmatrix}$$

Consider a new test point $\vec{z} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$. What is the predicted value $H(\vec{z})$?

Solution: In kernel regression, the prediction for a new point $\vec{z}$ is given by the sum of the dual weights multiplied by the kernel evaluations between the training points and the new point:

$$H(\vec{z}) = \sum_{i=1}^{3} \alpha_i \kappa(\vec{x}_i, \vec{z})$$

Step 1: Compute the kernel values $\kappa(\vec{x}_i, \vec{z})$:

- $\kappa(\vec{x}_1, \vec{z}) = (1 \cdot 2) + (2 \cdot 1) = 2 + 2 = 4$
- $\kappa(\vec{x}_2, \vec{z}) = (2 \cdot 2) + (0 \cdot 1) = 4 + 0 = 4$
- $\kappa(\vec{x}_3, \vec{z}) = (0 \cdot 2) + (3 \cdot 1) = 0 + 3 = 3$

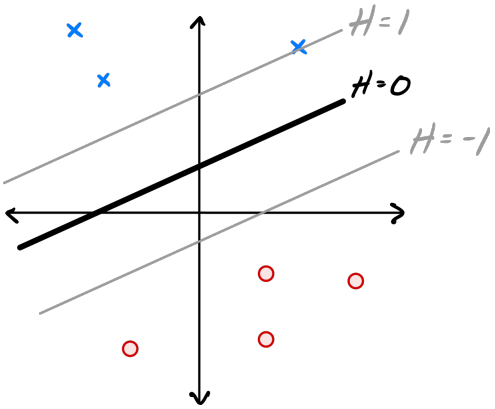
Step 2: Calculate the final prediction:

$$H(\vec{z}) = \alpha_1 \kappa(\vec{x}_1, \vec{z}) + \alpha_2 \kappa(\vec{x}_2, \vec{z}) + \alpha_3 \kappa(\vec{x}_3, \vec{z})$$
$$H(\vec{z}) = (1.0 \cdot 4) + (-0.5 \cdot 4) + (2.0 \cdot 3)$$
$$H(\vec{z}) = 4 - 2 + 6$$
$$H(\vec{z}) = 8$$

The prediction $H(\vec{z})$ is **8**.

Problem 5. (1 point)

The image below shows a linear prediction function H along with a data set; the "×" points have label +1 while the "○" points have label -1. Also shown are the places where the output of H is 0, 1, and -1.



True or False: Is it possible that the H shown in this figure was learned by training a Hard-SVM on this data set.

- ◯ True
- ● False

Solution: False.

The solution to the Hard-SVM problem is a hyperplane that separates the two classes with the maximum margin.

In this solution, the margin is not maximized: There is room for the "exclusion zone" between the two classes to grow, and for the margin to increase.

Problem 6. (2 points)

Which of the following functions are convex? Choose all that apply.

- ◯ $\log(x)$
- ● $|x|$ (absolute value of x)
- ◯ $\sin(x)$
- ● $4x^2 - 4x + 1$
- ● e^x
- ● $x^2 + \sin(x)$

Problem 7. (2 point)

Let $\mathcal{X} = \{(\vec{x}^{(1)}, y_1), \ldots, (\vec{x}^{(n)}, y_n)\}$ be a binary classification dataset with $\vec{x}^{(i)} \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Let $\mathcal{X}'$ be the dataset obtained by adding a single new feature to each feature vector in $\mathcal{X}$, keeping the first d features the same. Suppose a nearest neighbor classifier is trained on $\mathcal{X}$ with $k = 3$, achieving a training accuracy of 80%. Next, a nearest neighbor classifier is trained on $\mathcal{X}'$ with $k = 3$.

True or False: the new training accuracy is at least 80%.

- ● False
- ◯ True

Problem 8. (4 points)

Suppose we want to fit a ridge regression model to a data set. We could do this in three different ways:

1. Use the original features $\vec{x}$ directly. (Linear predictor with a regularization term)
2. First transform the inputs using a feature map $\vec{\phi}(\vec{x})$, then fit a linear model with a regularization term.
3. Use kernel ridge regression with a kernel $\kappa(\vec{x}, \vec{z})$ that corresponds to our feature space.

Conceptually explain how these three approaches are related. Make sure to include answers to these questions:

a) (2 points) Conceptually, explain how these three approaches are related.

Solution: In linear ridge regression, we are learning to minimize

$$\arg\min_{\vec{w}} \|Z\vec{w} - y\| + \lambda\|w\|$$

where

$$Z = \begin{cases} X, & \text{if using the original features,} \\ \Phi\Phi^T, & \text{if using feature maps,} \\ K, & \text{if using kernel ridge regression.} \end{cases}$$

. This unifies the three ways of performing the ridge regression, but really we are just doing ridge regression in either the original data space or the feature space.

b) (1 point) Why might we prefer approach 2 over approach 1?

Solution: The benefit of feature maps is the allow us to model the data nonlinearly if we choose nonlinear features, giving us more power/flexibility to model our data.

c) (1 point) Why might we prefer approach 3 over approach 2?

Solution: The benefit of using the kernel trick (i.e. kernel ridge regression) is that instead of actually computing each feature vector for each sample, the kernel trick allows us to save on time/computation by just computing the kernel function between the two samples.

Problem 9. (6 points)

Consider a linear classifier with initial weight vector $\vec{w}^{(0)} = [1, 1]^T$ and a small dataset $\mathcal{D}$ consisting of two points in $\mathbb{R}^2$:

- $(\vec{x}_1, y_1) = ([2, 0]^T, 1)$
- $(\vec{x}_2, y_2) = ([0, 2]^T, -1)$

a) (1 points) Calculate the average Perceptron loss.

Solution: The Perceptron loss for a single point is defined as $L_P = \max(0, -y(\vec{w}^T\vec{x}))$. First compute the scores:

$$\vec{w}^{(0)T}\vec{x}_1 = 2, \qquad \vec{w}^{(0)T}\vec{x}_2 = 2.$$

For the dataset $\mathcal{D}$:

- $L_1 = \max(0, -(1)(2)) = 0$
- $L_2 = \max(0, -(-1)(2)) = 2$

Average Loss: $\frac{0+2}{2} = 1$

b) (1 points) Calculate the average Hinge loss.

Solution: The Hinge loss is defined as $L_H = \max(0, 1 - y(\vec{w}^T\vec{x}))$.

- $L_1 = \max(0, 1 - (1)(2)) = \max(0, -1) = 0$
- $L_2 = \max(0, 1 - (-1)(2)) = \max(0, 3) = 3$

Average Loss: $\frac{0+3}{2} = 1.5$

c) (1 points) Calculate the average Square loss.

Solution: The Square loss is defined as $L_S = (y - \vec{w}^T \vec{x})^2$.

- $L_1 = (1-2)^2 = (-1)^2 = 1$
- $L_2 = (-1-2)^2 = (-3)^2 = 9$

Average Loss: $\frac{1+9}{2} = 5$

d) (3 points) Using the Hinge loss and a learning rate of $\eta = 0.5$, perform one step of Gradient Descent. Report the updated weight vector $\vec{w}^{(1)}$. *(Hint: For the gradient of the max function, assume the subgradient $\nabla \max(0, f(w)) = 0$ if $f(w) < 0$ and $\nabla f(w)$ otherwise.)*

Solution: Given learning rate $\eta = 0.5$, the weight update is $\vec{w}^{(1)} = \vec{w}^{(0)} - \eta \nabla_{\vec{w}} L_{avg}$.
The gradient of the Hinge loss for a single point is:

$$\nabla_{\vec{w}} L = \begin{cases} -y\vec{x} & \text{if } y(\vec{w}^T \vec{x}) < 1 \\ 0 & \text{otherwise} \end{cases}$$

- For $(\vec{x}_1, y_1)$: $y_1(\vec{w}^T \vec{x}_1) = 2 \geq 1 \implies \nabla_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$
- For $(\vec{x}_2, y_2)$: $y_2(\vec{w}^T \vec{x}_2) = -2 < 1 \implies \nabla_2 = -(-1) \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$

Average Gradient:

$$\nabla_{avg} = \frac{1}{2} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 2 \end{bmatrix} \right) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Update Step:

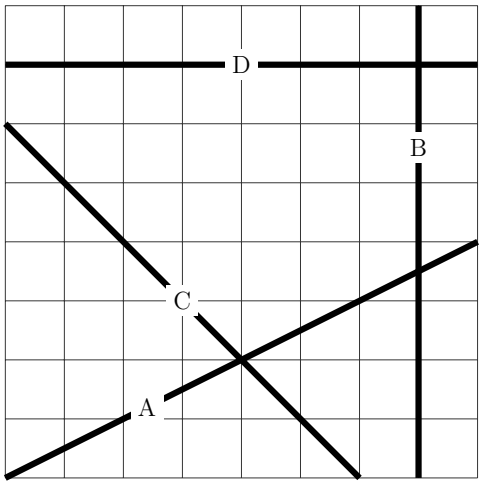
$$\vec{w}^{(1)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0.5 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 - 0.5 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$

Problem 10. (2 point)

Suppose a linear prediction function $H(\vec{x}) = w_0 + w_1 x_1 + w_2 x_2$ is trained to perform classification, and the weight vector is found to be $\vec{w} = (1, 2, 2)^T$.

The figure below shows four possible decision boundaries: A, B, C, and D. Which of them is the decision boundary of the prediction function H?

You may assume that each grid square is 1 unit on each side, however, we do not know where the origin is. You can assume that x_1 is the horizontal axis and x_2 is the vertical axis.



- ◯ D
- ◯ A
- ● C
- ◯ B

Solution: It must be C.

There are a couple of ways of solving this. First, you might remember from lecture (and the homework, where we proved the property) that the vector (w_1, w_2) is orthogonal to the decision boundary. In this case, this means that the boundary is orthogonal to $(2, 2)^T$, which is a vector pointing straight up and to the right. Boundary C is the only boundary that is orthogonal to this vector.

Another way to do this without recalling the property above is to substitute the weight vector into the equation of a linear prediction function:

$$\begin{aligned} H(\vec{x}) &= w_0 + w_1 x_1 + w_2 x_2 \\ &= 1 + 2 \cdot x_1 + 2 \cdot x_2 \\ &= 1 + 2x_1 + 2x_2. \end{aligned}$$

If you recall that the decision boundary is the set of points where $H(\vec{x}) = 0$, then you can solve for where $H = 0$ to find $x_2 = (-2x_1 - 1)/2$. This is the equation of a line with slope -1, and C is the only line with slope -1.

Before turning in your exam, please check that your name is on every page.