DS-GA 1003 Spring 2026 Midterm Review

Overview: This document is a collection of practice problems and details about the midterm for DS-GA 1003 (Spring 2026). If you are studying for the midterm, making sure you can do all of these problems and all the problems on the problem sets *only with the help of an* 8.5×11 *double-sided cheatsheet* should prepare you for the midterm.

Midterm Details

Basic Information. The midterm for this class will take place on **March 10, 2026 2:45pm - 4:45pm** during the usual class time, in the usual place in lieu of the lecture. *Please arrive early if you can, because we will be starting at 2:45pm sharp.* You will be allowed the following materials:

- A pen or pencil. I will provide sheets of scrap paper during the midterm.
- One-double sided 8.5×11 sheet of paper containing whatever materials you like. This can be handwritten or typed and there is no restriction on what this sheet contains.

Any collaboration between students or with outside sources via phones, laptops, "smart glasses," smoke signals, pagers, carrier pigeons, etc. is *strictly prohibited*.

Midterm Format. You can be assured that the format of the midterm will be as follows:

- One section for each of the first six lectures of the course (up to and including *MLE & Conditional Probability Models*).
- Each section except the *MLE & Conditional Probability Models* section will have exactly two parts: three True/False questions and one multi-part short answer problem, worth a total of 18 points.
  - **True/False.** Each True/False problem will be worth 3 points each. You will need to answer whether a statement is true or false with a one-sentence justification. 1 point is awarded for the correct True/False without justification, but the full 3 points will only be awarded for a correct justification.
  - **Short answer.** Each short answer problem is worth 9 total points. The short answer problems will involve working through a short toy machine learning problem, similar to the homework but easier. The samples in this document should be a good representation of their difficulty level.
- The *MLE & Conditional Probability Models* section of the exam will only have a single short answer problem worth 9 points.

Section 1: Statistical Learning Framework

Topics to Understand

1. In the supervised statistical learning framework, what is a *hypothesis*, *input space*, *output space*, and *action space*?
2. Understand the assumption of having an underlying data distribution over the input space and output space.
3. What is the *risk* of a hypothesis, and how do we measure that with respect to a data-generating distribution?
4. What is the *Bayes hypothesis*? What is the *Bayes risk*? Can you calculate these quantities if given a simple data-generating distribution? Can you calculate these quantities for the zero-one loss and the squared loss?
5. What is the *empirical risk* of a hypothesis? What is the empirical risk minimizer? What does this have to do with the finite dataset we are given in a machine learning problem? How does this relate to the *risk*?
6. What is a *hypothesis class*? What is meant by the complexity of this class? Be able to give a few examples of hypothesis classes.
7. What are the "three types of error" that guide our study of machine learning: *estimation error*, *approximation error*, and *optimization error*? Make sure to understand how these quantities change when we vary different parts of the learning problem, i.e. the size or complexity of the hypothesis class, the size of our dataset, our method of optimization, etc.
8. What is meant by *representation*, *generalization*, and *optimization*, and how do they relate to the errors above?
9. How does one run *polynomial regression* just using the machinery we developed for linear regression (Problem Set 1)?
10. Understand and be able to state the setup for linear regression under squared loss (in both scalar and matrix-vector form), and know how to derive the empirical risk minimizer of that problem by taking gradients and optimizing via calculus (Problem Set 1).

Sample True/False Problems

The exam itself will only include three True/False problems, but we have included several more for you to practice with.

Directions. For the following True/False problems, clearly state whether the statement is True/False and give a short justification (each needs at most one sentence). Any statement evaluated as True *must always* be True; if there exists *any* counterexample to the statement, you should mark it as False.

True/False Problem 1.1

The *approximation error* in the statistical learning framework is always positive or zero.

Solution

True. The approximation error is defined as $R(h^*_{\mathcal{H}}) - R(h^*)$, where h^* is the Bayes optimal hypothesis and $h^*_{\mathcal{H}}$ is the best hypothesis in $\mathcal{H}$. The Bayes optimal hypothesis is the best hypothesis over *all functions* for this distribution, while $\mathcal{H}$ is a subset of the class of all functions, so $R(h^*)$ must always be smaller.

True/False Problem 1.2

Suppose that we have a pair of hypothesis classes $\mathcal{H}_1$ and $\mathcal{H}_2$, such that $\mathcal{H}_1 \subset \mathcal{H}_2$. Then, the approximation error under $\mathcal{H}_1$ is *always* less than or equal to the approximation error of $\mathcal{H}_2$.

Solution

False. If $\mathcal{H}_1 \subset \mathcal{H}_2$, then $\mathcal{H}_2$ contains all functions in $\mathcal{H}_1$ and potentially better functions. Therefore, the best function in $\mathcal{H}_2$ will have risk equal to or lower than the best in $\mathcal{H}_1$.

True/False Problem 1.3

The *estimation error* is in the statistical learning framework can be a negative quantity.

Solution

False. The estimation error is defined as $R(\hat{h}) - R(h^*_{\mathcal{H}})$. Since $h^*_{\mathcal{H}}$ minimizes the true risk over $\mathcal{H}$, the empirical risk minimizer cannot have lower true risk than h^* (even if it has minimal *empirical risk*).

True/False Problem 1.4

One must know the true data-generating distribution in order to compute the estimation error.

Solution

True. In order to compute the estimation error, we need $R(\hat{h})$ and the minimum true risk $R(h^*_{\mathcal{H}})$, which both require expectations over the data-generating distribution.

True/False Problem 1.5

One must know the true data-generating distribution in order to compute the empirical risk.

Solution

False. Empirical risk is calculated solely using the finite observed dataset, so the true distribution is not needed to compute this.

True/False Problem 1.6

Viewing the dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ as a set of n random variables drawn from the joint distribution $P_{\mathcal{X}\times\mathcal{Y}}$, the optimization error when using full batch gradient descent is a random variable.

Solution

True. Since the dataset is drawn randomly from a distribution, the dataset itself is a random variable. Consequently, the hypothesis learned via gradient descent (and its resulting optimization error) is also a random variable.

Sample Short Answer Problems

The exam itself will only include one multi-part short answer problem, but we have included a couple for you to practice with.

Short Answer Problem 1.1. Consider the following binary classification problem. For the class $y = 0$, x is sampled from $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ with equal probability; for the class $y = 1$, x is sampled from $\{7, 8, 9, 10\}$ with equal probability. Assume that both classes are equally likely. For this problem, we consider the zero-one loss:

$$\ell(\hat{y}, y) := \mathbf{1}\{\hat{y} \neq y\} = \begin{cases} 1 & \text{if } \hat{y} \neq y \\ 0 & \text{if } \hat{y} = y \end{cases}$$

1. (2 points) Consider the constant predictor $f(x) = 1$. Under this distribution and the zero-one loss, what is the risk of this predictor?

2. (4 points) What is the Bayes hypothesis for this distribution, under the zero-one loss function? This can be written in terms of a a function $f^* : \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \to \{0, 1\}$. What is the Bayes risk, $R(f^*)$ under the zero-one loss?

3. (3 points) Suppose that you are using the hypothesis class of all step functions with a single jump, defined as any function parameterized by $\theta, c_1, c_2 \in \mathbb{R}$ with the form:

$$h_{\theta,c_1,c_2}(x) = \begin{cases} c_1 \text{ if } x < \theta \\ c_2 \text{ if } x \geq \theta \end{cases}$$

Given the dataset of six points $\{(1, 0), (2, 0), (3, 0), (7, 0), (7, 1), (8, 1)\}$, what is a possible empirical risk minimizer $\hat{h}$ for this dataset with respect to this hypothesis class and the zero-one loss? What is the empirical risk of $\hat{h}$ with respect to this dataset?

Solution

1. Under the zero-one loss, the risk is $R(f) = \mathbb{E}[\mathbf{1}\{f(x) \neq y\}] = \Pr(f(x) \neq y)$. Plugging in the constant predictor, we have $\Pr(y \neq f(x)) = \Pr(y \neq 1) = \Pr(y = 0) = 0.5$ because both classes are equally likely.

2. Recall from class that $f^*(x) = \arg\max_{a \in \{0,1\}} \Pr(y = a \mid x)$. We can construct f^* as follows:

   - For $x \in \{1, \ldots, 6\}$, we know $\Pr(y = 0 \mid x) = 1$, so we assign $f^*(x) = 0$.
   - For $x = 10$, we know that $\Pr(y = 1 \mid x) = 1$, so we assign $f^*(x) = 1$.
   - For $x \in \{7, 8, 9\}$, we recall Bayes' rule:

$$\Pr(y \mid x) = \frac{\Pr(x \mid y) \Pr(y)}{\Pr(x)} \propto \frac{1}{2} \cdot \Pr(x \mid y).$$

     Therefore, we can compare the likelihoods $\Pr(x \mid y = 0) = 1/9$ and $\Pr(x \mid y = 1) = 1/4$. Because $\Pr(x \mid y = 1)$ is greater, we assign $f^*(x) = 1$.

   Finally, to compute the Bayes risk

$$R(f^*) = \Pr(f^*(x) \neq y),$$

   we note that errors only occur when predicting $y = 1$ for $x \in \{7, 8, 9\}$ but the actual label is $y = 0$. The probability of this event is:

$$R(f^*) = \Pr(x \in \{7, 8, 9\}, y = 0) = \Pr(y = 0) \cdot \Pr(x \in \{7, 8, 9\} \mid y = 0) = 0.5 \cdot \frac{3}{9} = 1/6.$$

3. Split the step function at $\theta = 3.5$, so our ERM step function is $f(x) = \mathbf{1}\{x \geq 3.5\}$. Empirical risk is $1/6$, because it must misclassify the point at $x = 7$.

Short Answer Problem 1.2. Consider the following regression problem. Let $\mathcal{X} = [-5, 5]$, $\mathcal{Y} = \mathcal{A} = \mathbb{R}$, and suppose the data-generating distribution is specified by the marginal distribution $P_{\mathcal{X}}$ where $x \sim \text{Unif}[-5, 5]$ and the conditional distribution is $P_{\mathcal{Y}|X}$, where for each $x \in \mathcal{X}$, we have $y \sim N(2 + x, 1)$ (the Normal distribution with mean $2 + x$ and variance 1). For this problem, suppose we draw the following dataset of five points from this distribution:

$$D = \{(0, 2), (0, 1.5), (1, 3.1), (3, 4.7), (-3, 0)\}$$

Throughout this problem, we will be evaluated on the squared loss

$$\ell(\hat{y}, y) = (\hat{y} - y)^2.$$

1. (2 points) What is the Bayes hypothesis for this problem? What is the Bayes risk?

2. (3 points) Suppose you can use the hypothesis space of all possible functions $f : [-5, 5] \to \mathbb{R}$ for this problem. In this case, what is an empirical risk minimizer for this dataset, and what is the corresponding minimum empirical risk?

3. (2 points) Suppose you now consider using the hypothesis class of affine functions $\mathcal{H}_1 = \{x \mapsto wx + w_0 : w, w_0 \in \mathbb{R}\}$. Write down a design matrix X for this problem and a vector of outputs y such that the solution to $\hat{w} = (X^\top X)^{-1} X^\top y$ would give you the empirical risk minimizer for the above dataset D with respect to this hypothesis class. You only need to write down X and y but you do not need to solve for $\hat{w}$ or calculate the empirical risk.

4. (2 points) Suppose you now consider using the hypothesis class of degree two polynomials $\mathcal{H}_2 = \{x \mapsto w_2 x^2 + w_1 x + w_0 : w_0, w_1, w_2 \in \mathbb{R}\}$. Write down a design matrix X for this problem and a vector of outputs y such that the solution to $\hat{w} = (X^\top X)^{-1} X^\top y$ would give you the empirical risk minimizer for the above dataset D with respect to this hypothesis class.

Solution

1. Recall that, under the squared loss, the Bayes hypothesis is $h^*(x) = \mathbb{E}[y \mid x]$. In this case, $y \sim N(2+x, 1)$, so $\mathbb{E}[y \mid x] = 2 + x$. To compute the Bayes risk, we have

$$R(f^*) = \mathbb{E}[(f^*(x) - y)^2] = \mathbb{E}[(\mathbb{E}[y \mid x] - y)^2],$$

which is the conditional variance. By the distribution above, this is 1, so $R(f^*) = 1$.

2. Since the hypothesis space is all possible functions, for $x \in \{1, 3, -3\}$, we can choose $\hat{f}(1) = 3.1, \hat{f}(3) = 4.7$ and $\hat{f}(-3) = 0$. To decide what $\hat{f}(0)$ should be we need to choose the best $\hat{y} \in \mathbb{R}$ solving the minimization problem:

$$\hat{y} \in \arg\min_{\hat{y}} \ (\hat{y} - 2)^2 + (\hat{y} - 1.5)^2.$$

By standard calculus, we get $\hat{y} = 7/4$, so we set $\hat{f}(0) = 7/4$. Therefore, a possible ERM is

$$\hat{f}(x) = \begin{cases} 7/4 & \text{if } x = 0 \\ 3.1 & \text{if } x = 1 \\ 4.7 & \text{if } x = 3 \\ 0 & \text{if } x = -3 \\ 0 & \text{otherwise} \end{cases}$$

3. We include columns for x and the bias w_0:

$$X = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 1 \\ 3 & 1 \\ -3 & 1 \end{bmatrix} \quad y = \begin{bmatrix} 2 \\ 1.5 \\ 3.1 \\ 4.7 \\ 0 \end{bmatrix}.$$

4. We include columns for x^2, x, and the bias w_0:

$$X = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 9 & 3 & 1 \\ 9 & -3 & 1 \end{bmatrix} \quad y = \begin{bmatrix} 2 \\ 1.5 \\ 3.1 \\ 4.7 \\ 0 \end{bmatrix}.$$

Section 2: Optimization and Gradient Descent

Topics to Understand

1. Understand the statement of the closed form solution for linear regression: if $X \in \mathbb{R}^{n \times d}$ with $n \geq d$ and $\text{rank}(X) = d$, then the closed form solution is $\hat{w} = (X^\top X)^{-1} X^\top y$.

2. Know the *gradient descent algorithm* for unconstrained optimization. Understand that this can be used for any unconstrained optimization problem where you can take the derivative of an objective function and be prepared to use it for a differentiable problem you haven't seen before.

3. Be sure to know how to do gradient descent on a toy problem (i.e. write down what the update step should be, take a few updates by hand, etc.)

4. Understand what a *step size* is for gradient descent.

5. Understand the *"rough derivation"* of gradient descent in the lecture notes using the idea of *linear approximation*.

6. Understand the statement of the *descent lemma* and what it implies about various properties of gradient descent (i.e. what step size to choose, what the norm of the gradient tells us, what it says about local vs. global minima).

7. Understand the statement of *GD on convex, smooth functions* and understand why convexity helps gradient descent (just at a high level, as we didn't prove this).

8. Understand what the condition of *smoothness* entails and how to check if a function is smooth.

9. Understand the difference between *stochastic gradient descent* and gradient descent. Understand how to do stochastic GD if given a toy problem.

10. Understand how *minibatch gradient descent* improves upon stochastic gradient descent.

Sample True/False Problems

The exam itself will only include three True/False problems, but we have included several more for you to practice with.

Directions. For the following True/False problems, clearly state whether the statement is True/False and give a short justification (each needs at most one sentence). Any statement evaluated as True *must always* be True; if there exists *any* counterexample to the statement, you should mark it as False.

True/False Problem 2.1

Suppose you have a convex, twice-differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ whose Hessian is given by $2I$, where I is the $d \times d$ identity matrix. Then, we are *guaranteed* that gradient descent with $\eta = 1/4$ converges to a *global minimum*.

Solution

True. The eigenvalues of the Hessian are 2. The maximum step size that is stable by the descent lemma is $2/L$, where $L = 2$ in this case. $\eta = 1/4$ is within the allowable step size of $\eta \leq 1$.

True/False Problem 2.2

Suppose you have a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ whose Hessian is given by $3I$, where I is the $d \times d$ identity matrix. Then, we are *guaranteed* that gradient descent with $\eta = 1/4$ converges to a *global minimum*.

Solution

True. The function is convex because the Hessian is positive definite, and $L = 3$. Plugging into descent lemma, we have $\eta \leq 2/3$ are the stable values, and $\eta = 1/4$ satisfies this.

True/False Problem 2.3

Consider the function $f(w) = \sin w$, and consider running gradient descent on this function starting at $w_0 = 0$. With a step size $\eta = 1/2$, we are *guaranteed* that gradient descent will converge to a *global minimum*.

Solution

True. Although $\sin w$ is a nonconvex function with many local minima, the local minima all have the same value of -1. Starting at $w_0 = 0$ and running gradient descent will get us to the local minimum at $w = -\pi/2$.

True/False Problem 2.4

Suppose we are trying to minimize the function $f(w) = -w^2$. Running gradient descent with step size $\eta = 1/2$ for this function is guaranteed to converge to a *global minimum*.

Solution

False. The function $f(w) = -w^2$ is unbounded below (no global minimizer), so GD cannot converge to a global minimum.

True/False Problem 2.5

The gradient of a function $F : \mathbb{R}^d \to \mathbb{R}$ points in the direction where F decreases the fastest.

Solution

False. ∇F points in the direction of steepest *increase*. Its negative, $-\nabla F$, points in the direction of steepest *decrease*.

True/False Problem 2.6

Suppose we want to employ *early stopping*, which is the technique of just stopping gradient descent at a pre-specified step count decided before doing any optimization. Early stopping is a strategy to decrease *optimization error*.

Solution

False. Early stopping is typically used to reduce overfitting/generalization error, *not* optimization error. It will increase optimization error typically because running an optimization algorithm for more steps typically decreases the objective function value.
This problem was slightly vaguely worded — if we know that we are doing GD on a convex function, then we can definitively say that we are hurting the optimization error.

Sample Short Answer Problems

The exam itself will only include one multi-part short answer problem, but we have included a couple for you to practice with.

Short Answer Problem 2.1. In this problem, we consider a two-dimensional least squares linear regression problem where $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \mathcal{A} = \mathbb{R}$, with the twist that we are concerned about doing *weighted* linear regression. In this case, we are given a dataset and weights $\{(x^{(i)}, y^{(i)}, \gamma^{(i)})\}_{i=1}^n$ and we need to minimize the *weighted empirical risk*:

$$\hat{R}_n(w) = \sum_{i=1}^{n} \gamma^{(i)} (w^\top x^{(i)} - y^{(i)})^2.$$

where $\gamma^{(i)} > 0$ are positive weights given to each example.

1. (5 points) Write the objective function $\hat{R}_n(w)$ in matrix-vector form. That is, you should write $\hat{R}_n(w)$ as a function of a matrix $X \in \mathbb{R}^{n \times d}$, a vector $w \in \mathbb{R}^d$, a vector $y \in \mathbb{R}^n$, and a matrix of weights $G \in \mathbb{R}^{n \times n}$. Make sure you specify how $\gamma^{(i)}, w, x^{(i)}$, and $y^{(i)}$ all figure into these matrices and vectors.

2. (4 points) Write the gradient descent update step for moving from $w^{(t)}$ to $w^{(t+1)}$ in matrix-vector form (i.e. using the matrices and vectors outlined in the previous sub-problem). Partial credit is awarded if you can write out the step in scalar form (but only matrix-vector form will be awarded full credit).

Solution

1. Let $G \in \mathbb{R}^{n \times n}$ be the diagnoal matrix with $G_{i,i} = \gamma^{(i)}$. Then, we can rewrite the objective as
$$\hat{R}_n(w) = (Xw - y)^\top G(Xw - y).$$
Alternatively, taking $G^{1/2}$ as the diagonal matrix where $G_{i,i}^{1/2} = \sqrt{\gamma^{(i)}}$, we can also write this as
$$\hat{R}_n(w) = \|G^{1/2}(Xw - y)\|_2^2.$$

2. The gradient step is:
$$w^{(t+1)} \leftarrow w^{(t)} - 2\eta X^\top G(Xw^{(t)} - y).$$

Short Answer Problem 2.2. In this problem, we are interested in optimizing a logistic regression problem using gradient descent. We are in the binary classification setting, where $\mathcal{Y} = \{0, 1\}$, $\mathcal{A} = [0, 1]$, and our hypothesis class is the class of all linear functions with a bias term, composed with the sigmoid function:
$$h(x) = \sigma(w^\top x + w_0) \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$
The empirical risk minimization problem we are interested in solving is:
$$F(w, w_0) = \frac{1}{n} \sum_{i=1}^{n} -y^{(i)} \log(h(x^{(i)})) - (1 - y^{(i)}) \log(1 - h(x^{(i)})).$$

1. (2 points) Write the gradient descent update rule for $w^{(t)}$ and $w_0^{(t)}$ (this can be done separately).

2. (4 points) Suppose you are given the dataset of four points
$$((1, 0), 1), ((0, 1), 1), ((-1, 0), 0), ((0, -1), 0),$$
where each $x^{(i)} \in \mathbb{R}^2$. Starting from $w^{(0)} = (0, 0)$ and $w_0^{(0)} = 0$, take one step of gradient descent with $\eta = 1$. Compute by hand $w^{(1)}$ and $w_0^{(1)}$ (the parameters after a single step).

3. (3 points) After the single step of gradient descent from the previous problem, you decide to terminate the algorithm. You decide to threshold the probabilities given by $h(x) \in [0, 1]$ at $1/2$, predicting $\hat{y}^{(i)} = 1$ if $h(x) \geq 1/2$ and $\hat{y}^{(i)} = 0$ if $h(x) < 1/2$. For the dataset and discovered parameters in the previous problem, what is your classification on each of the four data points? What is your empirical risk with respect to the *zero-one loss*?

Solution

1. The gradient descent updates are given by:

$$w^{(t+1)} \leftarrow w^{(t)} - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} (h(x^{(i)}) - y^{(i)}) x^{(i)} \quad w_0^{(t+1)} \leftarrow w_0^{(t)} - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} (h(x^{(i)}) - y^{(i)}),$$

   where $h(x) = \sigma(w^\top x + w_0)$.

2. At $w = 0$ and $w^{(0)} = 0$, we have $h(x) = 0$ for all the points. We can compute the gradients:

$$\nabla F = (-0.25, -0.25) \quad \nabla_{w_0} F = 0.$$

   Therefore, taking the gradient step, we obtain $w^{(1)} = (0.25, 0.25)$ and $w_0^{(1)} = 0$.

3. We can determine that the scores $w^\top x + w_0$ are, in order for these examples, $0.25, 0.25, -0.25$ and -0.25. Therefore, the probabilities are > 0.5, > 0.5, < 0.5, and < 0.5, and the classifications are $\hat{y} = 1, \hat{y} = 1, \hat{y} = 0$, and $\hat{y} = 0$. All classifications are correct, so under zero-one loss, we have zero empirical risk.

Section 3: Regularization & Loss Functions

Topics to Understand

1. What is meant by the *complexity* of a hypothesis class? Can you give several example hypothesis classes and the corresponding measure of complexity?

2. How does the *complexity* of the hypothesis class influence the three main errors we study: optimization error, approximation error, and estimation error?

3. Understand polynomial regression and how the phenomenon of *overfitting* is exhibited in this example.

4. Understand model selection as applied to polynomial regression.

5. Understand the difference between *penalized ERM* and *constrained ERM* in the context of regularization.

6. Understand the *ridge (ℓ_2-regularized) regression* objective in the context of regularization. Be comfortable with the two forms (penalized and constrained) and how the regularization parameter (in lecture, we called this λ) affects the optimization problem.

7. Understand the analytic solution to the *ridge regression* problem and contrast it to the analytic solution to unregularized least squares.

8. Understand the *lasso (ℓ_1-regularized) regression* objective in the context of regularization. Be comfortable with the two forms (penalized and constrained) and how the regularization parameter (in lecture, we called this λ) affects the optimization problem.

9. Understand what is meant by the comparison of the "regularization paths" in ridge and lasso.

10. Understand, intuitively, when one might choose lasso over ridge (and vice versa) for a particular regression problem.

11. Understand the difference between squared loss and *absolute loss* for regression.

12. Understand what a *margin-based loss* is and how this relates to the binary classification problem.

13. Understand the relationship between *zero-one loss*, *hinge loss*, *perceptron loss*, *logistic loss*, and *square loss* in the context of binary classification. Be able to map these losses to their corresponding optimization problems.

Sample True/False Problems

The exam itself will only include three True/False problems, but we have included several more for you to practice with.

Directions. For the following True/False problems, clearly state whether the statement is True/False and give a short justification (each needs at most one sentence). Any statement evaluated as True *must always* be True; if there exists *any* counterexample to the statement, you should mark it as False.

True/False Problem 3.1

Recall the *ridge regression (ℓ_2-regularized)* objective in *penalized form*:

$$\sum_{i=1}^{n}(w^\top x^{(i)} - y^{(i)})^2 + \lambda\|w\|_2^2.$$

Focus on w_1, the first entry of w. By increasing λ, the entry w_1 will always *increase* in magnitude.

Solution

False. The norm of w generally decreases with larger λ. With larger and larger λ, ridge shrinks the coefficients to 0.

True/False Problem 3.2

You are given a regression problem with $d = 100$ features, and your task is to fit a linear model (i.e. you use the hypothesis class $\{x \mapsto w^\top x + w_0 : w \in \mathbb{R}^d, w_0 \in \mathbb{R}\}$). Your boss tells you that she suspects only a small number (≤ 10) of the features are relevant to the prediction problem, but she is having trouble figuring out which. Using *ridge regression*, you are easily able to return to your boss a parameter vector w^* whose entries are zero for 90 of the features.

Solution

False. Ridge (ℓ_2) regression tends to shrink the weights but does not enforce sparsity. Lasso (ℓ_1) regression is required for sparsity.

True/False Problem 3.3

Suppose you have a linear regression problem where the number of examples $n = 1000$ and the number of features $d = 1,000,000$. We can solve *penalized ridge regression* with $\lambda > 0$ such that $\hat{w} = (X^\top X + \lambda I)^{-1} X^\top y$ where $X \in \mathbb{R}^{n \times d}$ is the design matrix, $y \in \mathbb{R}^n$ is the vector of outputs, and $I \in \mathbb{R}^{d \times d}$ is the identity matrix.

Solution

True. The matrix $X^\top X + \lambda I$ is always invertible for $\lambda > 0$ because adding λI makes the matrix positive definite, regardless of the rank of X.

True/False Problem 3.4

Suppose we have a dataset $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$ and we attempt to run empirical risk minimization with the *logistic loss* on this dataset. If we find some hypothesis h such that the empirical risk with the logistic loss is exactly zero ($\hat{R}_n(h) = 0$), then the empirical risk of h under the *zero-one loss* is exactly zero as well.

Solution

True. Logistic loss is nonnegative and is only equal to 0 if every point is classified with probability 1 for its label, implying zero zero-one loss.

True/False Problem 3.5

Consider a binary classification problem. Under the *margin formulation* of *zero-one loss*, the classification on a datapoint (x, y) where $h(x) = 4$ and $y = -1$ has the zero-one loss evaluate to 0.

Solution

False. The margin is $yh(x) = -4$. This is negative, which means that we have a misclassfiication, and, hence zero-one loss is 1.

True/False Problem 3.6

Consider a dataset $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, and we want to minimize the empirical risk minimization problem over the class of linear functions (functions of the form $h_w(x) = w^\top x$) with respect to the zero-one loss. Applying gradient descent to this problem will output a minimizer w^* that has minimal zero-one loss on this dataset.

Solution

False. We cannot optimize zero-one loss with gradient descent because it is nondifferentiable.

Sample Short Answer Problems

The exam itself will only include one multi-part short answer problem, but we have included a couple for you to practice with.

Short Answer Problem 3.1. Consider the regularized linear regression problem with squared loss, in matrix-vector form:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|Xw - y\|_2^2 + \lambda R(w),$$

where $R(w)$ is a regularizer $R : \mathbb{R}^d \to \mathbb{R}$ and $X \in \mathbb{R}^{n \times d}$ is your design matrix of training examples and $y \in \mathbb{R}^n$ is your vector of outputs. Assume that the columns of X are orthonormal: $X^\top X = I$, where $I \in \mathbb{R}^{d \times d}$ is the identity matrix. Define:

$$z := X^\top y \in \mathbb{R}^d.$$

1. (3 points) Let $R(w) = \frac{1}{2}\|w\|_2^2$. Derive the minimizer w_{ridge} as a function of z and λ.
2. (3 points) Let $R(w) = \|w\|_1$. Show that the optimization decouples coordinate-wise, i.e. we can write the objective as a sum of d terms where each term only has w_j and z_j for $j \in \{1, \ldots, d\}$.
3. (3 points) Let $d = 3$, $z = (2, 0.6, -1.2)$, and $\lambda = 1$. Compute the ridge minimizer w_{ridge} for this problem.

Solution

1. If $R(w) = \frac{1}{2}\|w\|_2^2$, then the objective is

$$\frac{1}{2}\|Xw - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2.$$

   Using $X^\top X = I$, the minimizer satisfies $(I + \lambda I)w = z$, so we have

$$w_{\text{ridge}} = \frac{1}{1+\lambda}z.$$

2. If $R(w) = \|w\|_1$, then we can expand the the first part as

$$\frac{1}{2}\|Xw - y\|_2^2 = \frac{1}{2}(w^\top w - 2z^\top w + \|y\|_2^2) = \sum_{j=1}^{d} \frac{1}{2}w_j^2 - z_j w_j + y_j^2.$$

   Therefore, the full objective, when we add $\|w\|_1 = |w_1| + \cdots + |w_d|$, becomes:

$$\frac{1}{2}\|Xw-y\|_2^2+\lambda\|w\|_1 = \sum_{j=1}^{d} \frac{1}{2}w_j^2 - z_j w_j + \lambda|w_j| + y_j^2 = \|y\|_2^2 + \sum_{j=1}^{d} \frac{1}{2}w_j^2 - z_j w_j + \lambda|w_j|.$$

   But the $\|y\|_2^2$ term is a constant and does not affect the minimization. We see that the sum $\sum_{j=1}^{d} \frac{1}{2}w_j^2 - z_j w_j + \lambda|w_j|$ decouples over w_j and z_j.

3. With $d = 3$, $z = (2, 0.6, -1.2)$, and $\lambda = 1$, we get

$$w_{\text{ridge}} = \frac{1}{2}(2, 0.6, -1.2) = (1, 0.3, -0.6).$$

Short Answer Problem 3.2. You want to find the parameters $w \in \mathbb{R}^2$ for a linear hypothesis $h(x) = w^\top x$. You are given two datasets of two examples each:

$$D_1 = \{((1,0), 2), ((1,1), -1)\}$$
$$D_2 = \{((1,1), +1), ((1,-1), -1)\}$$

1. (2 points) Write down a loss function and a one-sentence justification for why that loss function may be appropriate for dataset D_1.

2. (2 points) You are told that, though dataset D_2 only has several examples, there will be more examples coming soon. Regardless, you are told that every example is labeled with one of two outcomes. Name a loss function that is appropriate for dataset D_2 and explain why you needed the information that every example is labeled with one or two outcomes first before deciding that loss function.

3. (5 points) Suppose that you are now evaluating a linear hypothesis you already found with parameters $w = (1, -1)$. For both D_1 and D_2, write down the empirical risk of w on the data above using the loss functions you chose in (1) and (2).

Solution

1. D_1 is a regression problem, so a reasonable loss is $\ell(\hat{y}, y) = (\hat{y} - y)^2$, the squared loss.

2. We can use the logistic loss $\ell(m) = \log(1 + \exp(-m))$ where $m = y\hat{y}$, the margin. We needed to know that the labels are binary because otherwise it would be ambiguous whether this was a regression or classification problem, just as stated.

3. On D_1, the predictions are $h(1, 0) = 1$ and $h(1, 1) = 0$. The squared losses are $(1-2)^2 = 1$ and $(0-(-1))^2 = 1$. Therefore, the empirical risk is equal to 1 (divide by 2 because $n = 2$).

   On D_2, the predictions are $h(1, 1) = 0$ and $h(1, -1) = 2$. The losses are therefore $\log 2$ and $\log(1 + e^2)$ and therefore the empirical risk is $\frac{1}{2}(\log 2 + \log(1 + e^2))$.

One could also have chosen absolute loss and hinge loss for this problem, for instance.

Section 4: Convex Optimization & SVM

Topics to Understand

1. Understand that the central property for why convex optimization is nice is that all local minima are global minima (as a corollary, gradient descent is guaranteed to find a global minima for unconstrained convex optimization problems).

2. Understand the definition of *convex set* and *convex function*.

3. Understand how to apply the definition of *convex function* (as well as its *first-order* and *second-order* conditions) to multivariate functions (by, e.g., taking gradients and Hessians).

4. Understand the basic form of a *convex optimization problem* and how to convert any convex optimization problem into the standard form (all $\leq$ inequality constraints).

5. Understand how to convert a constrained convex optimization problem into its *Lagrangian*.

6. Understand *duality*, *weak duality*, and *strong duality* and how these relate the primal and dual optimization problems.

7. Understand the *Lagrangian dual function* associated with a given optimization problem, and how it relates to the original optimization problem.

8. Understand the *complementary slackness conditions* and the Slater's condition given in class to ensure complementary slackness holds (proofs not needed).

9. Understand *margin* and *hinge loss* in the context of the SVM optimization problem and how it relates to the ξ_i variables (the "slack" variables).

10. Understand the derivation of the *soft-margin SVM* optimization problem.

11. Understand the geometric derivation of the *hard-margin SVM* optimization problem (discussed in lab).

12. Understand the basic geometry of SVM and what is meant by finding the hyperplane of *maximum margin*.

13. Given a toy SVM classification problem, understand geometrically where the SVM hyperplane should go and how removing points may affect the hyperplane.

14. Understand what a *support vector* is, both from the optimization and geometric perspective of the SVM.

15. Know the *dual form* of the SVM optimization problem, and what the Lagrange multipliers correspond to.

16. Understand the relationship between the Lagrange multipliers and the form of w^* in the solution to the SVM optimization problem (and how this relates to *support vectors* and *margin*).

Sample True/False Problems

Directions. For the following True/False problems, clearly state whether the statement is True/False and give a short justification (each needs at most one sentence). Any statement evaluated as True *must always* be True; if there exists *any* counterexample to the statement, you should mark it as False.

True/False Problem 4.1

The function $f(x) = \ln(x + 3)$ is a convex function.

Solution

False. This function $\ln(x + 3)$ is concave on $x > -3$, not convex.

True/False Problem 4.2

The function $F : \mathbb{R}^2 \to \mathbb{R}$ defined by $F(x_1, x_2) = x_1^2 + x_2^2$ is a convex function.

Solution

True. One can take the Hessian, which is $2I$, and see that it is positive definite.

True/False Problem 4.3

Every convex function has a *unique* global minimum.

Solution

False. Convex functions can have multiple minimizers. For instance, $f(x) = 1$ is convex and every x is a minimizer.

True/False Problem 4.4

Suppose we have a binary classification problem of $n = 1000$ examples with $d = 100$ features. We want to run the *soft-margin SVM* on this problem. On day one, we formulate the soft-margin SVM problem on this dataset and run a convex optimization solver on it. The next day, you get 1000 more examples, so you now have a binary classification problem of $n = 2000$ examples and $d = 100$ features. You run SVM again. The number of constraints in your SVM problem are now strictly less than on day one.

Solution

False. Soft-margin SVM has a constraint for each example. Therefore, increasing n increases (not decreases) the number of constraints.

True/False Problem 4.5

Suppose you switch the regularizer in SVM to the ℓ_1 norm, $\|w\|_1 = |w_1| + \cdots + |w_d|$. The *soft-margin SVM* is still a convex optimization problem.

Solution

True. Norms are convex, and $\|w\|_1$ is a norm. The hinge loss is also convex, so adding $\|w\|_1$ to the hinge loss in the objective of SVM maintains convexity.

Sample Short Answer Problems

The exam itself will only include one multi-part short answer problem, but we have included a couple for you to practice with.

Short Answer Problem 4.1. Consider the following $d = 1$ soft-margin SVM objective:

$$\min_{w,b\in\mathbb{R}} C\sum_{i=1}^{3}\max\{0, 1 - y^{(i)}(wx^{(i)} + b)\} + \frac{1}{2}w^2.$$

Consider the toy dataset of three examples:

$$(x^{(1)}, y^{(1)}) = (-1, -1) \quad (x^{(2)}, y^{(2)}) = (1, +1) \quad (x^{(3)}, y^{(3)}) = (2, +1).$$

Set $C = 1$ for this problem.

1. (2 points) Explain in one or two sentences why the objective is convex.

2. (3 points) Evaluate the objective with the two candidate solutions:

$$w = 1, b = 0 \quad \text{(candidate A)}$$
$$w = \frac{1}{2}, b = 0 \quad \text{(candidate B)}$$

   Which candidate has a better objective value?

3. (3 points) For the better candidate, identify which points are (i) correctly classified but inside the margin, (ii) misclassified, or (iii) on/outside the margin.

4. (1 point) Noting that the input space in this problem is $\mathcal{X} = \mathbb{R}$ and the output space is $\mathcal{Y} = \{-1, +1\}$, draw the points and the better candidate on the real line, where points labeled -1 are represented by X's and points labeled $+1$ are represented by O's.

Solution

1. The objective function is a sum of the hinge loss (which is convex) and the ℓ_2-regularizer/norm (which is convex). A sum of convex functions is convex.

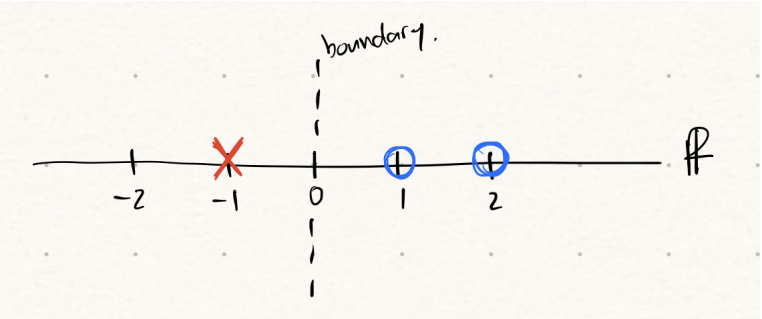
2. We can evaluate the candidates just by checking the empirical risks:

   - Candidate A ($w = 1$, $b = 0$)
     - Point $(-1, -1)$ has margin $-1 \cdot -1 = 1$, and loss $\max(0, 1 - 1) = 0$.
     - Point $(1, 1)$ has margin $1 \cdot 1 = 1$, and loss $\max(0, 1 - 1) = 0$.
     - Point $(2, 1)$ has margin $2 \cdot 1 = 2$, and loss $\max(0, 1 - 2) = 0$.
     - The regularizer has value $\frac{1}{2}\|w\|^2 = 1/2$.
     - Therefore, the total objective value is $1/2$.
   - Candidate B ($w = 1/2$, $b = 0$).
     - Point $(-1, -1)$ has margin $-1 \cdot -0.5 = 0.5$, and loss $\max(0, 1 - 1) = 0.5$.
     - Point $(1, 1)$ has margin $1 \cdot 0.5 = 0.5$, and loss $\max(0, 1 - 0.5) = 0.5$.
     - Point $(2, 1)$ has margin $1 \cdot 1 = 1$, and loss $\max(0, 1 - 1) = 0$.
     - The regularizer has value $\frac{1}{2}\|w\|^2 = 1/2(0.25) = 0.125$.
     - Therefore, the total objective value is 1.125.

   Therefore, Candidate A is better because its objective value is lower.

3. The first two points $(-1, -1)$ and $(1, +1)$ are correct and on the margin, with $y(wx + b) = 1$. These are support vectors. The third point $(2, +1)$ is correct but outside the margin, with $y(wx + b) > 1$, so it is not a support vector.

4. The decision boundary is when $wx + b = 0$, which occurs at $x = 0$. This is a drawing:



Short Answer Problem 4.2. In this problem, we consider the SVM objective, as described in lecture and lab.

1. (3 points) Recall the SVM objective is equivalent to minimizing the average hinge loss

with ℓ_2 regularization. What if we minimize the hinge loss without regularization? Assuming the data are linearly separable, why is the solution not unique without regularization?

2. (3 points) Consider the following dataset:

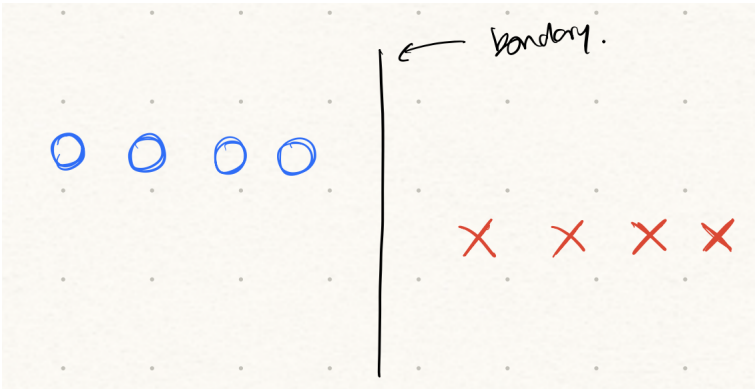
where the O's are positive examples with label $+1$ and the X's are negative examples with label -1. Draw the decision boundary given by a linear (hard-margin) SVM trained on this dataset, and circle the support vectors.

3. (3 points) Draw another decision boundary for this dataset that achieves zero empirical risk. Explain, in the context of this decision boundary, why SVM is preferable for finding the linear decision boundary for this problem.

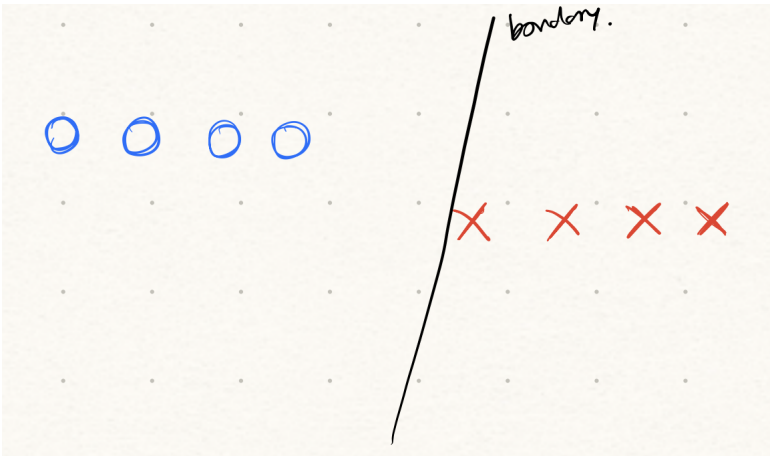
Solution

1. If the data is separable, we can scale w and b by an arbitrary factor. This increases the margin and drives hinge loss to 0. But without regularization to penalize the magnitude of w, we can always increase the parameters w, so the solution is not unique.

2. A drawing would look like this:



3. A drawing would look like this:



The SVM is preferable to this boundary because it finds the decision boundary of *maximum margin*. This is more stable/robust and tends to generalize better than arbitrary zero-error separators.

Section 5: Features & Kernels

Topics to Understand

1. Understand what is meant by *feature extraction* and how this is typically one of the first steps before even doing any machine learning.
2. Understand that mapping a binary classification problem that is not linearly separable to a higher-dimensional space can make it linearly separable.
3. Understand some motivations for developing "smarter features" for a machine learning problem.
4. Understand the motivations for kernel methods: they mainly alleviate memory and computational costs involved with computing in a higher-dimensional feature space. They give us access to higher-dimensional feature spaces without extra *computational* cost.
5. Understand why the polynomial kernels and quadratic feature maps shown in class have an inner product that can be computed *only with* inner products in the original d-dimensional space.
6. Understand why kernels make computation faster in the previous examples.
7. Understand the definition of a *positive semidefinite matrix* and how it relates to developing kernels through *Mercer's Theorem*.
8. Be able to identify whether a function $k(x, z)$ is a kernel.
9. Understand the "high-level recipe" for applying kernel methods and how computing a *kernel matrix* factors into this.
10. Understand the statement of the *representer theorem* and the types of objective functions it applies to.
11. Understand why SVM and ridge regression are two cases that the representer theorem applies to just by looking at the form of the objective function.

Sample True/False Problems

True/False Problem 5.1

Suppose you are dealing with a binary classification problem on $\mathcal{X} = \{0,1\}^5$ and $\mathcal{Y} = \{-1,+1\}$. You also know that the true distribution is such that, for any $x \in \mathcal{X}$, there exists $c^* : \mathcal{X} \to \mathcal{Y}$ such that $y = c^*(x)$ (there is a deterministic function mapping all inputs to outputs). Then, *there exists* a kernel $\Phi\colon \mathcal{X} \to \mathbb{R}^d$ for $d > 5$ such that the data is linearly separable.

Solution

True. Since $\mathcal{X} = \{0,1\}^5$ is finite with $|\mathcal{X}| = 32$ distinct inputs and the labeling is deterministic (no conflicting labels), we can construct a one-hot feature map into $\mathbb{R}^{32}$ (and $32 > 5$) that sends each distinct input to a distinct orthogonal unit vector, making any binary labeling linearly separable.

True/False Problem 5.2

Consider the objective function:

$$F(w) = \sum_{i=1}^{n} |w^\top x^{(i)} - y^{(i)}| + \lambda \|w\|^2,$$

where $\lambda > 0$. By the Representer Theorem, the optimal weight vector $w^* \in \mathbb{R}^d$ that minimizes this objective can be written as $w^* = \sum_{i=1}^n \alpha_i x^{(i)}$ for some $\alpha \in \mathbb{R}^n$.

Solution

True. The objective has the form $L(w^\top x^{(1)}, \ldots, w^\top x^{(n)}) + R(\|w\|)$ where $R(\|w\|) = \lambda\|w\|^2$ is a strictly increasing function of $\|w\|$, so the Representer Theorem applies and guarantees that the optimal w^* lies in $\text{span}\{x^{(1)}, \ldots, x^{(n)}\}$.

True/False Problem 5.3

Mapping a dataset that is not linearly separable in its original lower-dimensional space into a higher-dimensional feature space will *always* make it linearly separable.

Solution

False. If two data points share the same input but have different labels, any feature map sends them to the same point, so no hyperplane can ever separate them regardless of how high-dimensional the feature space is.

True/False Problem 5.4

A machine learning method can be *kernelized* as long as every feature vector in both the optimization problem and the prediction function appears exclusively within an inner product with another feature vector.

Solution

True. This is exactly the condition needed to apply the kernel trick: if $\phi(x)$ only ever appears as $\phi(x^{(i)})^\top \phi(x^{(j)})$, we can replace each such inner product with a kernel function $k(x^{(i)}, x^{(j)})$ and never need to compute ϕ explicitly.

True/False Problem 5.5

Suppose you train an SVM on a dataset of $n = 300,000$ points with $d = 10$ original features using the RBF kernel. The resulting kernel matrix K you compute will have dimensions 10×10.

Solution

False. The kernel matrix K has one entry per pair of training points, so its dimensions are $n \times n = 300{,}000 \times 300{,}000$; the number of original features $d = 10$ does not determine the size of the kernel matrix.

Sample Short Answer Problems

Short Answer Problem 5.1. In this problem, we consider a one-dimensional binary classification problem (where the input space is $\mathcal{X} = \mathbb{R}$ and the output space is $\mathcal{Y} = \{-1, +1\}$) with the following toy dataset of three examples:

$$(x^{(1)}, y^{(1)}) = (-2, -1) \quad (x^{(2)}, y^{(2)}) = (2, -1) \quad (x^{(3)}, y^{(3)}) = (0, +1).$$

1. (2 points) Is the dataset linearly separable in the original 1D input space $\mathcal{X} = \mathbb{R}$? Briefly explain why or why not in one or two sentences.

2. (4 points) Consider the feature map

$$\phi(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}.$$

Map the three training examples into the new 2D feature space. Is the data linearly separable in the new space? If so, provide a weight vector w and a bias w_0 that correctly separates the classes with *strictly positive margin* on every point.

3. (3 points) Write down the explicit kernel function $k(x, x') = \phi(x)^\top \phi(x)$ corresponding to this feature map. Compute the 3×3 kernel matrix K for this dataset.

Solution

1. **No.** The dataset is not linearly separable in the 1D input space $\mathcal{X} = \mathbb{R}$. The positive point ($x = 0$) is situated strictly between the two negative points ($x = -2$ and $x = 2$). A 1D linear classifier is a single point (a threshold) on the number line, which can only divide the line into a left half and a right half.

2. Mapping the points using $\phi(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$:

   - $\phi(x^{(1)}) = \phi(-2) = \begin{bmatrix} -2 \\ 4 \end{bmatrix} \quad (y^{(1)} = -1)$
   - $\phi(x^{(2)}) = \phi(2) = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \quad (y^{(2)} = -1)$
   - $\phi(x^{(3)}) = \phi(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (y^{(3)} = +1)$

   Yes, the data is linearly separable in this new 2D space. The positive point is at the origin, and the negative points are elevated at the coordinate $x_2 = 4$. A horizontal line can separate them.

   One valid separating hyperplane is defined by the weight vector $w = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$ and bias $w_0 = 2$.

3. The explicit kernel function is:

$$k(x, x') = \phi(x)^\top \phi(x') = \begin{bmatrix} x & x^2 \end{bmatrix} \begin{bmatrix} x' \\ (x')^2 \end{bmatrix} = xx' + x^2(x')^2$$

   The Kernel matrix $K \in \mathbb{R}^{3\times 3}$ has entries $K_{ij} = k(x^{(i)}, x^{(j)})$.

$$K = \begin{bmatrix} 20 & 12 & 0 \\ 12 & 20 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Short Answer Problem 5.2. You are given a training dataset of n examples $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ where $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$. You want to perform ridge regression in a high-dimensional feature space defined by feature map $\phi(x) \in \mathbb{R}^D$ where $D \gg d$. Recall that the optimization

problem for this is

$$\min_{w \in \mathbb{R}^D} F(w) = \sum_{i=1}^{n} (w^\top \phi(x^{(i)}) - y^{(i)})^2 + \lambda \|w\|^2.$$

where $\lambda > 0$.

1. (3 points) Suppose $x \in \mathbb{R}^2$ and you want to use the polynomial kernel $k(x, z) = (x^\top z + 1)^2$. Explicitly write out the corresponding feature map $\phi(x)$ such that $k(x, z) = \phi(x)^\top \phi(z)$.

2. (2 points) The representer theorem guarantees that the optimal weight vector w^* lines in the span of the training examples in the higher-dimensional feature space. Let $K \in \mathbb{R}^{n \times n}$ be the kernel matrix for this dataset and let $\alpha \in \mathbb{R}^n$ be the coefficients of w in the span of the higher-dimensional inputs. Write the penalty term $\|w^*\|^2$ strictly in terms of α and the kernel matrix K.

3. (4 points) Substitute $w^* = \sum_{i=1}^{n} \alpha_i \phi(x^{(i)})$ into the objective $F(w)$ to write the objective entirely in terms of the vector α, the vector of labels $y \in \mathbb{R}^n$, the kernel matrix K, and $\lambda > 0$. Then, take the gradient of the new objective with respect to α and solve for the optimal dual variables α^* directly via calculus.

Solution

1. Let $x = [x_1, x_2]^\top$ and $z = [z_1, z_2]^\top$. We expand the kernel function:

$$\begin{aligned} k(x, z) &= (x_1 z_1 + x_2 z_2 + 1)^2 \\ &= (x_1 z_1)^2 + (x_2 z_2)^2 + 1^2 + 2(x_1 z_1)(x_2 z_2) + 2(x_1 z_1)(1) + 2(x_2 z_2)(1) \\ &= x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1 x_2 z_1 z_2 + 2x_1 z_1 + 2x_2 z_2 + 1 \end{aligned}$$

   By grouping the x and z terms to form a dot product $\phi(x)^\top \phi(z)$, we define the explicit feature map:

$$\phi(x) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1 x_2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ 1 \end{bmatrix}$$

2. Using the Representer Theorem substitution $w^* = \sum_{i=1}^n \alpha_i \phi(x^{(i)})$:

$$\begin{aligned} \|w^*\|^2 &= (w^*)^\top w^* \\ &= \left(\sum_{i=1}^n \alpha_i \phi(x^{(i)}) \right)^\top \left(\sum_{j=1}^n \alpha_j \phi(x^{(j)}) \right) \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \phi(x^{(i)})^\top \phi(x^{(j)}) \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K_{ij} \end{aligned}$$

   In matrix notation, this quadratic form is exactly:

$$\|w^*\|^2 = \alpha^\top K \alpha$$

   (This was already shown in class in lecture).

Solution

3. Substituting this and the result from Part 2 into the primal objective yields the dual objective in terms of α:

$$F(\alpha) = (y - K\alpha)^\top (y - K\alpha) + \lambda \alpha^\top K \alpha$$

Expanding the objective (and using the fact that $K = K^\top$ because kernel matrices are symmetric):

$$\begin{aligned} F(\alpha) &= y^\top y - 2y^\top K\alpha + \alpha^\top K^\top K\alpha + \lambda \alpha^\top K\alpha \\ &= y^\top y - 2y^\top K\alpha + \alpha^\top (K^2 + \lambda K)\alpha \end{aligned}$$

Now, take the gradient with respect to the vector α and set it to the zero vector:

$$\nabla_\alpha J(\alpha) = -2Ky + 2(K^2 + \lambda K)\alpha = 0$$

Divide by 2 and factor out K:

$$K(K\alpha + \lambda\alpha) = Ky$$

$$K(K + \lambda I)\alpha = Ky$$

Assuming K is positive definite/invertible, we can multiply both sides by K^{-1}:

$$(K + \lambda I)\alpha = y$$

Solving for α^* requires multiplying by the inverse of $(K + \lambda I)$, which is guaranteed to be invertible for $\lambda > 0$:

$$\alpha^* = (K + \lambda I)^{-1} y$$

Section 6: Probabilistic Models & MLE

Topics to Understand

Sample True/False Problems

Problem 6.1

Suppose your data is $z_1, \ldots, z_n$ drawn from some unknown distribution. Suppose your likelihood function is $L(\theta) := \Pr[z_1, \ldots, z_n \mid \theta]$, where $\theta \in \Theta$ is the parameter vector for some family of parameters Θ. Then, we finding the MLE, if it exists, is equivalent to *maximizing* $\log L(\theta)$ with respect to θ.

Problem 6.2

Performing MLE for the regression problem where $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$ under the assumption that the conditional distribution given x is Gaussian with mean $\mu = \theta^\top x$ and variance σ^2 is equivalent to ERM on the squared loss over the class of linear models.

Problem 6.3

From the probabilistic modeling perspective, logistic regression can be seen as assuming that the conditional distribution given x is a Bernoulli distribution with success parameter $\sigma(\theta^\top x)$ where $\theta \in \mathbb{R}^d$ is a vector and $\sigma(z) = (1 + e^{-z})^{-1}$ is the sigmoid function.