

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.

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 .

True/False Problem 1.3

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

True/False Problem 1.4

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

True/False Problem 1.5

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

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.

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 function $f^* : \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \rightarrow \{0, 1\}$. 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?

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 $f : [-5, 5] \rightarrow \mathbb{R}$ 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 $\mathcal{H}_2 = \{x \mapsto w_2x^2 + w_1x + w_0 : w_0, w_1, w_2 \in \mathbb{R}\}$. 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.

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.

True/False Problem 2.1

Suppose you have a convex, twice-differentiable function $F : \mathbb{R}^d \rightarrow \mathbb{R}$ 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*.

True/False Problem 2.2

Suppose you have a differentiable function $F : \mathbb{R}^d \rightarrow \mathbb{R}$ 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*.

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*.

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*.

True/False Problem 2.5

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

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*.

True/False Problem 2.7

Increasing the minibatch size for *minibatch gradient descent* always results in a gradient step

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 subproblem). Partial credit is awarded if you can write out the step in scalar form (but only matrix-vector form will be awarded full credit).

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:

$$h(x) = \sigma(w^\top x + w_0) \quad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

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*?

Section 3: Regularization & Loss Functions

Topics to Understand

Sample True/False Problems

Sample Short Answer Problems

Section 4: Convex Optimization & SVM

Topics to Understand

Sample True/False Problems

Sample Short Answer Problems

Section 5: Features & Kernels

Topics to Understand

Sample True/False Problems

Sample Short Answer Problems

Section 6: Probabilistic Models & MLE

Topics to Understand

Sample True/False Problems