CS 281A - Midterm

October 28, 2025

This exam is open book and open notes. In answering these questions, quote any results, techniques or algorithms that you use precisely and provide appropriate citations. **You may not discuss the exam with anyone, whether they are in the class or not**. If you need clarification for a specific question please ask a private question on Ed, do not post publicly, and do not e-mail the course staff.

Since this exam is a take-home, please use the extra time to make your solutions as clear as possible. We will not give credit to solutions we cannot read. Include as much of your reasoning as possible. We are interested in seeing how you think about the problem, not only if you get the correct answer. However, do not be long winded for the sake of being long winded. Be clear, precise, and concise.

This exam is to be turned in on bCourses by **Wednesday, October 29th at 9:59am**. We will not accept late submissions.

1. Suppose we'd like to build a classification rule for a finite population Ω of pairs (x_i, y_i) where x_i are integers and y_i are assumed to be in $\{-1, 1\}$. Let n denote the size of Ω.

    1. Compute the function $\hat{y}_{\text{me}}(x)$ that minimizes the number of classification errors on this population.
    2. Suppose we instead choose to minimize the square-loss over all possible functions f:

    $$R_{\text{sq}}[f] = \frac{1}{2n} \sum_{i=1}^{n} (f(x_i) - y_i)^2$$

    Compute the function f_{sq} that minimizes this risk.

    3. Define the classification rule

    $$\hat{y}_{\text{sq}}(x) = \begin{cases} 1 & f_{\text{sq}}(x) \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

    Compute the fraction of times that $\hat{y}_{\text{sq}}$ makes an error. How does it compare to the error rate of $\hat{y}_{\text{me}}$?

2. Let $S = \{(x_i, y_i)\}$ be a set of example-label pairs with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Suppose we want to minimize the empirical loss

$$\frac{1}{2n} \sum_{i=1}^{n} (\langle w, x_i \rangle - y_i)^2$$

with respect to $w \in \mathbb{R}^d$. Assume the x_i span $\mathbb{R}^d$ and there is a solution $w_\star$ that achieves zero loss.

Analyze the performance of stochastic gradient descent at minimizing this loss. Assume that w is initialized equal to 0. At each iteration, a stochastic gradient is computed from a single example from S, sampled independently with replacement. Let (x_{i_k}, y_{i_k}) denote the example-label pair selected at iteration k. Assume the step size is a constant.

1. Show that the SGD iterates take the form

$$w_{k+1} = w_\star + A_k(w_k - w_\star)\,.$$

2. Compute $\mathbb{E}[A_k]$.
3. Show that there exists a step size such that

$$\| \mathbb{E}[w_T - w_\star] \| \leq \beta^T \|w_\star\|$$

for some $\beta < 1$.

3. You want to build a simple classifier on a d-dimensional space, which uses the rule

$$f(x) = \begin{cases} 1 & x_i \geq \alpha \text{ and } x_j \geq \beta \\ -1 & \text{otherwise} \end{cases}$$

That is, the classifier examines exactly two dimensions and declares the example to be positive if the values in those dimensions are both large enough.

You are given three sets that are guaranteed to be i.i.d. samples from a larger set S:

1. S_{train} with n_{train} data points.
2. S_{test} with n_{test} data points.
3. S_{val} with n_{val} data points.

You use the set S_{train} to find the best parameters α and β for each possible choice of dimensions i and j. This yields a set of predictors $\hat{f}_{ij}$. You select the $\hat{f}_{ij}$ that has the lowest error on S_{test}. Call the selected predictor $\hat{f}_{\text{best}}$.

Let E_{test} denote the 0-1 classification error of $\hat{f}_{\text{best}}$ on S_{test} and E_{val} denote the 0-1 classification error of $\hat{f}_{\text{best}}$ on S_{val}. Provide a bound on the probability that $|E_{\text{test}} - E_{\text{val}}|$ will be less than ϵ.

4. Consider a classification problem where you have access to K models that were trained by hyperparameter tuning on a train-test split $(S_{\text{train}}, S_{\text{test}})$. Here, the data sets are assumed to be representative examples for a classification problem. Suppose someone gives you a set S_{val} that they assert was generated to be identically distributed to S_{test}. Let $E_{\text{test}}[f_k]$ and $E_{\text{val}}[f_k]$ denote the error of f_k on S_{test} and S_{val} respectively.

    1. Make a hypothetical scatter plot of points that represents what you'd expect the relationship between E_{test} (x-axis) and E_{val} (y-axis) to look like. Explain your reasoning for why your plot looks the way it does.
    2. Describe conditions under which you'd expect $E_{\text{test}}[f_k]$ and $E_{\text{val}}[f_k]$ to be nearly equal.