Homework 1

Hugo Chardon
Department of Statistics, UC Berkeley

DUE DATE: Friday, March 6th, 2026 (11:00am)

Instructions:

- Questions marked with a cup of coffee (☕) are harder and are for students enrolled in Stat 254. Students from 154, don't hesitate to try! All efforts appreciated.
- Submit your answers as a single pdf on gradescope. You can use whatever tool you like to produce the pdf (LaTeX, markdown, quarto documents, Jupyter, scanned handwritten notes for mathematical problems (make them easy to read!), etc...).

1 Classification

In class, we addressed the problem of binary classification with 0/1 loss, $\ell(f(x), y) = \mathbf{1}(f(x) \neq y)$. This loss makes sense when mistakes are symmetric, meaning that the only thing that all wrong predictions should be equally penalized: predicting $f(x) = 1$ when the truth was $y = 0$ and predicting $f(x) = 0$ when the truth was $y = 1$.

In real applications, one mistake can be significantly worse than the other. This is in fact the way in which the theory of tests in statistics has been developed over the past century, with Type I and Type II errors. One way to enforce this in machine learning is to use the imbalanced binary loss.

Problem 1. Consider the binary prediction problem with label convention $\mathcal{Y} = \{0, 1\}$. Consider the following loss. Let $w \in (0, 1)$ and define the loss

$$\ell_w(\widehat{y}, y) = w\mathbf{1}(\widehat{y} = 1, y = 0) + (1 - w)\mathbf{1}(\widehat{y} = 0, y = 1)\,. \tag{1}$$

1. Give real-world examples of classification where the imbalanced loss (1) is relevant.
2. Check that for the weight $w = 1/2$, you recover the usual 0/1 loss that we discussed in class (see lecture notes). Recall the Bayes predictor for this loss.
3. Check that for a general $w \in (0, 1)$, the risk associated with the loss (1) is given, for any predictor f and random pair (X, Y), by

$$L^{(w)}(f) = w\mathbf{P}(f(X) = 1, Y = 0) + (1 - w)\mathbf{P}(f(X) = 0, Y = 1) \tag{2}$$

4. For the remaining questions, you may follow the steps of the proof of Proposition 2.1 in the lecture notes.

Show that the Bayes classifier is given by

$$f_{\mathsf{B}}(x) = \mathbf{1}(\eta(x) > w)\,, \quad \text{where } \eta(x) = \mathbf{P}(Y = 1|X = x) = \mathbf{E}[Y|X = x] \tag{3}$$

denotes the regression function as usual.

5. Show that the Bayes risk $L_{\mathsf{B}}^{(w)} = L^{(w)}(f_{\mathsf{B}})$ satisfies

$$L_{\mathsf{B}}^{(w)} = \mathbf{E}\big[\min\big\{(1-w)\eta(X), w(1-\eta(X))\big\}\big] \tag{4}$$

6. Show that the excess risk $L^{(w)}(f) - L_{\mathsf{B}}^{(w)}$ of any classifier f is given by

$$L^{(w)}(f) - L_{\mathsf{B}}^{(w)} = \mathbf{E}\big[|\eta(X) - w| \cdot \mathbf{1}(f(X) \neq f_{\mathsf{B}}(X))\big]\,.$$

Problem 2 (☕). Let $(X, Y) \in \mathcal{X} \times \{0, 1\}$ be a random pair whose distribution P is "zero-error". This means that X contains all the information about Y: if you know X, then you automatically know Y. We work with the standard 0/1 loss $\ell(\widehat{y}, y)$

1. Using the properties of conditional expectations, show that the "zero-error" property amounts to saying that $Y = \eta(X)$.

2. Under the "zero-error" property, show the following improvement of Proposition 2.2 in the lecture notes:

$$L(\widehat{f}_{\widehat{\eta}}) - L_{\mathsf{B}} \leqslant 4\mathbf{E}\big[(\widehat{\eta}(X) - \eta(X))^2\big]\,.$$

(This is in fact the conditional expectation with respect to the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, since $\widehat{\eta}$ is an estimator of the regression function η.)

3. Show that under the more general *margin* condition

$$|2\eta(X) - 1| \geqslant h \text{ a.s.},$$

then

$$L(\widehat{f}_{\widehat{\eta}}) - L_{\mathsf{B}} \leqslant \frac{4}{h}\mathbf{E}\big[(\widehat{\eta}(X) - \eta(X))^2\big]\,.$$

2 Linear modeling

Taken from R.Giordano's homework in previous iterations of the course

Problem 3 (An unhelpful colleague)**.** For this problem, we will consider the problem of predicting human bodyfat percentage from easyily measured body dimensions such as height, weight, abdomen circumference, (`Height`, `Weight`, and `Abdomen`) and so on.

Since measuring bodyfat directly requires expensive equipment (such as immersion in a water tank to measure body volume), it would be useful to be able to approximate bodyfat using more easily obtained measures. Assume that we have a dataset consisting of n observations, where the vector of covariates $X \in \mathbb{R}^d$ consists of easily–obtained body measurements, such as `Height`, `Weight`, and `Abdomen`. We denote by $X_i^{(j)}$ the j-th coordinate of X_i. For example, if $j = 1$ is the `Weight` variable, $X_i^{(1)}$ is the weight of person number i.

The corresponding $Y \in [0, 1]$ is an expensive–to–obtain but precise measure of bodyfat. You and a colleague would like to use X to accurately predict an unseen Y on a held–out dataset.

Unfortunately, your colleague has not taken a linear modeling class.

1. Your colleague notes that the regressors are on different scales, since they are measured in different units. For example, `Height` is measured in inches, and `Weight` in pounds. They suggest standardizing using US averages, replacing $X^{(j)}$ with

$$\overline{X}^{(j)} := \frac{X^{(j)} - \mu_j}{\sigma_j}$$

where μ_j and σ_j are the US average and standard deviation of regressor j (the variable $X^{(j)}$), and then regressing on $\overline{X}$ instead.

When they do so, they find that the test error does not change at all. Can you explain why (in detail)?

2. Chastened, your colleague suggests that maybe it's the difference between normalized height and weight that would helps us predict. After all, it makes sense that height should only matter relative to weight, and vice versa. So they run the regression on the pairwise differences, so they create a new vector of features Φ

$$\Phi := \overline{X}^{(j)} - \overline{X}^{(k)},$$

for all pairs (j, k) with $j > k$. That is, $\Phi \in \mathbb{R}^{d(d-1)/2}$, because there are $d(d-1)/2$ distinct pairs of normalized covariates. Unfortunately, when they regress Y on Φ using standard statistical software, they find that the fitted values do not change, and many of the coefficients are estimated as `NaN`. Can you explain why (in detail)?

3. Increasingly frustrated, your colleague suggests a research project where you improve your fit by regressing $Y \sim Z$ for new regressors Z of the form $Z = AX$, where the matrix A is optimized in an outer loop. That is, they suggest finding

$$Z(A) = AX\,,$$

use the (empirical) risk

$$\widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle \theta, Z_i(A)\rangle)^2\,,$$

and fitting

$$\widehat{\theta}_n(A) = \arg\min_{\theta} \widehat{L}_n(\theta)\,, \quad \widehat{A} = \arg\min_{A} \widehat{L}_n(\widehat{\theta}_n(A))\,.$$

They would then use the prediction $\widehat{Y} = \langle \widehat{\theta}_n(\widehat{A}), Z(A)\rangle$. Will this (complicated) procedure produce a better fit to the data than simply regressing Y on X? Why or why not?

Problem 4 (Incorrectly specified linear regression)**.** Assume the following model for the pair $(Z, X, Y) \in \mathbb{R}^p \times \mathbb{R}^d \times \mathbb{R}$:

$$Y = \langle \theta^*, X\rangle + \langle \gamma^*, Z\rangle + \varepsilon\,, \tag{5}$$

where ε is centered, has finite variance σ^2 and is independent of X and Z. We also assume that X and Z are centered.

1. Following the steps of the lecture notes, check that the Bayes predictor is

$$f_{\mathsf{B}}(x, z) = \mathbf{E}[Y|X = x, Z = z] = \langle \theta^*, X\rangle + \langle \gamma^*, Z\rangle\,.$$

We assume as usual that $X_1, \dots, X_n$ are i.i.d., $Z_1, \dots, Z_n$ are i.i.d., but the distribution of (Z, X, Y) is such that X and Z are possibly correlated (they don't have to be independent, just like in the usual linear model, the coordinates of X don't have to be independent). We define

$$\Sigma_X = \mathbf{E}[XX^\top], \quad \Sigma_Z = \mathbf{E}[ZZ^\top], \quad \Sigma_{X,Z} = \mathbf{E}[XZ^\top]. \tag{6}$$

We assume Σ_X and Σ_Z to be invertible. Note that, since X and Z are centered, these are actual covariance matrices. We consider the misspecified linear model where we regress Y only on X, that is, we consider the OLS estimator $\widehat{\beta}_n$ for the pairs $(X_1, Y_1), \dots, (X_n, Y_n)$.

2. Check that $\widehat{\beta}_n$ converges to some β^*, and give the expression of this limiting vector in terms of the other quantities of the problem.

3. Suppose we are using our linear regression for causal inference, and want to know θ^*, which we take as the causal effect of X on Y.

   Suppose $\Sigma_{X,Z} \neq 0$. Does $\beta^* = \theta^*$? Interpret this result in terms of unobserved confounders.

4. Suppose we are using linear regression for prediction, so we do not care about estimating θ^*, but rather care about the asymptotic quality of our estimates $\widehat{Y}^* = \langle \beta^*, X \rangle$. Write an expression for the expected squared limiting bias,

$$\mathbf{E}\big[(f_{\mathsf{B}}(X, Z) - \widehat{y}^*)^2\big]. \tag{7}$$

5. Compare (7) with the bias of the "correct" prediction based on X alone

$$\mathbf{E}\big[(f_{\mathsf{B}}(X, Z) - \langle \theta^*, X \rangle)^2\big]. \tag{8}$$

   In particular, show that the expected squared limiting bias of the "incorrect" regression is always smaller than the expected squared bias of the prediction $\langle \theta^*, X \rangle$.

6. Unobserved confounders are severely problematic for causal inference. Is the same true for prediction problems? Justify your answer in terms of the results from the previous questions.

Problem 5 (Taylor expansion invariance?)**.** Consider a regression problem with a pair $(X, Y) \in \mathbb{R} \times \mathbb{R}$ (that is, we use only one explanatory variable). Our class of predictor consists of Taylor series centered at some $x_0 \in \mathbb{R}$, meaning that we consider predictors of the form

$$f_\theta(x) = \sum_{j=0}^{K} \theta_j (x - x_0)^j, \quad \theta \in \mathbb{R}^K + 1, \quad K \geqslant 0. \tag{9}$$

Let $\widehat{\theta}_n$ denote the OLS estimator of the best population predictor θ^* (the one that minimizes the population risk for the square loss).

1. Show that the prediction $\widehat{Y} = \langle \widehat{\theta}_n, X \rangle$ does not depend on the choice of the point x_0. In this sense, it does not matter for linear prediction where you center your Taylor series.

2. Show that the previous conclusion is not true in general if you do not include a constant term in the regression (i.e., if you start the Taylor expansion (9) at $j = 1$ instead of $j = 0$).

3 Computational problem

Taken from N.Zhivotovskiy's homework in previous iterations of the course

For this last part you will need to perform computations on a real dataset. Use the standard machine learning libraries such as scikit learn in Python.

Problem 6 (prediction via LDA and QDA)**.** For this exercise we will use the abalone dataset from https://archive.ics.uci.edu/dataset/1/abalone to predict the sex from length and diameter.

1. Reduce the dataset to only those rows which have `sex = F` or `M`. Ignore the infants `I`. Also remove every column other than sex, length and diameter.
2. Divide the dataset into two parts, training and test. Ensure that their sizes are approximately equal.
3. Fit an LDA and a QDA model to the training set to predict sex from length and diameter.
4. Use your models to predict sex for the test dataset. Which model does better, and what metric did you use to measure that?
5. Perform two scatter plots, one for LDA and one for QDA, each containing both the training and test data (it is a 2D scatter plot). For the training data use two different colors (other than red) to indicate the class (`F` or `M`). Plot the test data similarly, but use a different shape. But, if a data point was predicted incorrectly, use red for its color instead.