CPSC 532D, Fall 2023: Assignment 1
due Monday, 18 September 2023, **12:00 noon**

Prepare your answers to these questions using LaTeX; hopefully you're reasonably familiar with it, but if not, try using Overleaf and looking around for tutorials online. Feel free to ask questions if you get stuck on things on Piazza (but remove any details about the actual answers to the questions…make a private post if that's tough). If you prefer, the `.tex` source for this file is available on the course website, and you can put your answers in `\begin{answer} My answer here... \end{answer}` environments to make them stand out if so; feel free to delete whatever boilerplate you want. Or answer in a fresh document.

Do assignment 1 alone; future ones will allow partners. If you look stuff up anywhere other than in SSBD or MRT, **cite your sources**: just say in the answer to that question where you looked. If you ask anyone else for help, **cite that too**. Please do not look at solution manuals / search for people proving the things we're trying to prove / etc. Also, please do *not* ask ChatGPT or similar models. It's okay to talk to others in the class about general strategies – if so, just say who and for which questions – but **not** to sit down and do the assignment together.

Submit your answers as a single PDF on Gradescope: here's the link. You'll be prompted to mark where each question is in your PDF; make sure you mark all relevant pages for each part (which save a surprising amount of grading time).

Please **put your name on the first page** as a backup, just in case. If something goes wrong, you can also email your assignment to me directly (`dsuth@cs.ubc.ca`).

1 Loss functions [55 points]

As a reminder, the general form of learning problems we'll usually work with in this course is as follows: $\mathcal{D}$ is some distribution over a space $\mathcal{Z}$, and $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ is a loss function.

For example, classification problems are often framed with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with the zero-one loss function $\ell(h, (x, y)) = \mathbb{1}(h(x) \neq y)$. The true risk is $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}} \ell(h, z)$, and the empirical risk is $L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$ for a sample $S = (z_1, \ldots, z_n) \sim \mathcal{D}^m$.[1]

(1.1) [5 points] Show that, for any given $h \in \mathcal{H}$, L_S is unbiased: $\mathbb{E}\, L_S(h) = L_{\mathcal{D}}(h)$.

Answer: TODO

(1.2) [5 points] Show that the expected zero-one loss for k-way classification ($\mathcal{Y} = [k] = \{1, \ldots, k\}$) is equal to one minus the expected accuracy (the portion of correct answers on samples from $\mathcal{D}$).

Answer: TODO

(1.3) [10 points] For the canonical ImageNet Large Scale Visual Recognition Challenge, images are given with one of a thousand possible labels, and one major way of evaluating those models is the top-5 accuracy: models can make 5 guesses at the label, and we count how often the correct label is one of those 5 guesses. Frame this in the language above: what kind of object does $h(x)$ output, and what does $\ell(h, (x, y))$ look like?

Answer: TODO

(1.4) [10 points] *Semantic segmentation* is a computer vision problem where we try to label each pixel of an image as belonging to one of k classes ("tree," "street," "dog," etc.). Let $S = ((x_1, y_1), \ldots, (x_n, y_n))$ where x_i are the given input images (in, say, $\mathbb{R}^{h \times w \times 3}$ and $y_i \in [k]^{h \times w}$ their corresponding pixel labels.[2] One typical evaluation metric is called mIoU ("mean intersection over union"). One minus the mIOU (to make a nicer "loss") is measured on a test set as follows:

$$Q_S = 1 - \frac{1}{k} \sum_{c=1}^{k} \frac{\text{\# of pixels from all images in } S \text{ that are \textit{correctly} predicted as } c}{\text{\# of pixels from all images in } S \text{ predicted as } c \text{ and/or with true label } c}.$$

Argue that this metric *cannot* be expressed using the form of loss function above on the given S. (A formal proof isn't necessary on this question, just a good intuitive argument.)

Answer: TODO

(1.5) [10 points] Recall the 1-nearest neighbour classifier: when trained on $S = ((x_1, y_1), \ldots, (x_m, y_m))$, the learned predictor $\hat{h}_S$ finds $\hat{i} \in \arg\min_{i \in [m]} \|x - x_i\|$ and then returns as its prediction $y_{\hat{i}}$. Can you write this algorithm as ERM? If so, give the loss function and hypothesis class and show it's an ERM; if not, argue why not.

Answer: TODO

(1.6) [10 points] Principal component analysis (PCA) is a common technique that can try to find an underlying low-dimensional structure by a linear mapping to a low-dimensional space: a data point $x \in \mathbb{R}^d$ is mapped to a latent code $z = Wx \in \mathbb{R}^k$, where $W \in \mathbb{R}^{k \times d}$ is a matrix with orthonormal rows ($WW^\top = I$) that we want to learn. To reconstruct a point from its latent code z, we take $W^\top z$. To

[1] The notation $\mathcal{D}^m$ here refers to a *product distribution*: a distribution which gives m independent and identically-distributed samples from $\mathcal{D}$.

[2] $[k]$ is semi-common notation for $\{1, 2, \ldots, k\}$; thus y_i is an $h \times w$ array of integers between 1 and k.

find W, we minimize the squared reconstruction error on a training set:

$$\operatorname*{arg\,min}_{W : WW^\top = I} \sum_{i=1}^{m} \|W^\top W x_i - x_i\|^2. \tag{PCA}$$

Frame PCA as an empirical risk minimization problem: what are the data domain $\mathcal{Z}$, the sample S, the hypothesis class $\mathcal{H}$, and the loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ such that the set of ERMs is exactly the set of solutions to (PCA)?

Answer: TODO

(1.7) [5 points] Frame the problem of fitting a Gaussian distribution to a set of independent scalar observations as loss minimization, like above: what are the data domain $\mathcal{Z}$, the sample S, the hypothesis class $\mathcal{H}$, and the loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ such that the ERM agrees with the maximum likelihood estimate?

Answer: TODO

2 Bayes optimality [35 points]

A Bayes-optimal predictor is a predictor which achieves the lowest possible error for any function, regardless of a choice of hypothesis class or anything like that.[3]

We'll consider loss functions of the form $\ell(h, (x, y)) = l(h(x), y)$, where $h : \mathcal{X} \to \hat{\mathcal{Y}}$ and $l : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$.[4] (We often have $\hat{\mathcal{Y}} = \mathcal{Y}$, as in binary classification, but not necessarily, as you may have seen in the previous question.)

A Bayes-optimal predictor has no pesky constraints on the form of function it's going to be, so it can just give an arbitrary different prediction for each x. Let $\mathcal{F}(x)$ denote the conditional distribution of y for a given x under $\mathcal{D}$: if $\mathcal{D}$ is deterministic, this won't be a very interesting distribution (a point mass), but in general it might be more complicated.

(2.1) [10 points] Argue that if h and g are predictors such that for every x, $\mathbb{E}_{y\sim\mathcal{F}(x)}\, l(h(x), y) \le \mathbb{E}_{y\sim\mathcal{F}(x)}\, l(g(x), y)$, then we necessarily have that $L_{\mathcal{D}}(h) \le L_{\mathcal{D}}(g)$.

Answer: TODO

Thus, we can find a generic Bayes-optimal predictor according to

$$f_{\mathcal{D},l}(x) \in \operatorname*{arg\,min}_{\hat{y}\in\hat{\mathcal{Y}}} \mathbb{E}_{y\sim\mathcal{F}(x)}\, l(\hat{y}, y).$$

(2.2) [5 points] Use the above formulation to argue that

$$f_{\mathcal{D},\text{0-1}}(x) = \begin{cases} 1 & \text{if } \Pr_{y\sim\mathcal{F}(x)}(y = 1) \ge \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

is Bayes-optimal for binary classification problems with 0-1 loss.

Answer: TODO

(2.3) [10 points] Use the above formulation to derive the Bayes-optimal predictor for a binary classification problem with the loss of an "is this mushroom edible" classifier:

$$l(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 0.01 & \text{if } \hat{y} = 0,\ y = 1 \\ 1 & \text{if } \hat{y} = 1,\ y = 0. \end{cases}$$

Answer: TODO

(2.4) [10 points] Use the above formulation to argue that

$$f_{\mathcal{D},\text{sq}}(x) = \mathbb{E}_{y\sim\mathcal{F}(x)}\, y$$

is Bayes-optimal for scalar regression problems with squared loss $l(\hat{y}, y) = (\hat{y} - y)^2$.

Answer: TODO

[3] As usual in this course, I'm ignoring issues of measurability and so on; this should all be formalizable by being appropriately careful and using "disintegrations" of probability measures, etc, but for the purpose of this question you can just ignore such issues.

[4] This is often how loss functions are defined in the first place; there are a few cases in the course where the more general ℓ form is more convenient (including in parts of Question 1), but for this question, the l form is a little easier.

3 Optimistic rates [10 challenge points]

Assignments in this course will generally have challenge questions. *These questions are harder than the other ones, and worth at most 10 points, so the effort:points ratio is far worse. If you never touch the challenge questions but get everything else right, you can still get a 90 (the lowest possible A+) in the course. But I think they're interesting questions, so if you have the time to spend, you might learn something.*

In this problem, assume that ℓ is an arbitrary loss bounded in $[0,1]$, and $\mathcal{H}$ is finite.

In the second lecture, we showed/will show *(depending on when you're reading this...)* the following bound on the statistical error of ERM:

$$\Pr\left(L_{\mathcal{D}}(\hat{h}_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(\mathcal{H}) \le \sqrt{\frac{2}{m} \log \frac{|\mathcal{H}|+1}{\delta}} \right) \ge 1 - \delta.$$

This $1/\sqrt{m}$ dependence is what's known as a "slow rate." In some settings, you can show a "fast rate" with $1/m$ dependence. (This gap is pretty big: if you observe 100 times as many samples, a $1/m$ rate will reduce the error by a factor of 100, while $1/\sqrt{m}$ would only reduce by 10.)

In previous years, I actually first proved a fast rate for finite hypothesis classes if you assume *realizability*: that there is some $h^* \in \mathcal{H}$ with $L_{\mathcal{D}}(h^*) = 0$. In that case, you can show a $\frac{1}{m} \log \frac{|\mathcal{H}|}{\delta}$ gap. (You can see the argument in Section 2.3.1 of the [SSBD] book, linked from the course site.)

One drawback of having this is that we have two totally separate analyses. If we know the problem is realizable, we get the nice $1/m$ rate. But as soon as $\min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) > 0$, we immediately jump up to the much worse rate.

We're going to prove an "optimistic" bound, one that smoothly interpolates between the two rates depending on the value of $L^* = \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$. This is going to take some more powerful machinery, and get a nastier bound, but the rate will be what we want.

One way to do this is based on *Bernstein's inequality*:

Proposition 3.1 (Bernstein, bounded variables)**.** *Let $X_1, \dots, X_m$ be independent random variables with means $\mu_i \in \mathbb{R}$, variances σ_i^2, and almost surely bounded in $[a,b]$. Then*

$$\Pr\left(\frac{1}{m} \sum_{i=1}^m (X_i - \mu_i) \ge \varepsilon \right) \le \exp\left(- \frac{m\varepsilon^2}{2\left(\frac{1}{m}\sum_{i=1}^m \sigma_i^2\right) + \frac{2}{3}(b-a)\varepsilon} \right). \tag{Bernstein}$$

(3.1) [4 points] Use Proposition 3.1 to show that for a fixed h, it holds with probability at least $1-\delta$ over the choice of $S \sim \mathcal{D}^m$ that

$$L_S(h) \le L_{\mathcal{D}}(h) + \frac{C_1 \log \frac{1}{\delta}}{m} + \sqrt{\frac{C_2 \log \frac{1}{\delta}}{m} L_{\mathcal{D}}(h)}. \tag{*}$$

for some (simple) universal constants C_1, C_2; give values for those constants.

Use this bound (don't prove it again) to show that with probability at least $1-\delta$,

$$L_S(h) \ge L_{\mathcal{D}}(h) - \frac{C_1 \log \frac{1}{\delta}}{m} - \sqrt{\frac{C_2 \log \frac{1}{\delta}}{m} L_{\mathcal{D}}(h)}. \tag{**}$$

You don't need to do this part to do the next one; you can just write that in terms of C_1 and C_2.

Hint: This is not *an exact inverse of the Bernstein probability bound; we're being a little loose here to get a simpler form.*

Answer: TODO

Now on to the bound. Let $\hat{h}_S$ denote an ERM, and let $h^* \in \arg\min_{h \in \mathcal{H}} L_\mathcal{D}(h)$, with loss $L^* = L_\mathcal{D}(h^*)$.[5]

(3.2) [6 points] Prove a bound on $L_\mathcal{D}(\hat{h}_S) - L^*$ in terms of L^*, $|\mathcal{H}|$, and m of the form

$$L_\mathcal{D}(\hat{h}_S) \le L^* + \mathcal{O}\left(\frac{1}{m} \log \frac{|\mathcal{H}| + 1}{\delta} + \sqrt{\frac{L^*}{m} \log \frac{|\mathcal{H}| + 1}{\delta}} \right).$$

For full credit, use explicit constants in your answer, not $\mathcal{O}$.

You can assume that $\frac{1}{m} \log \frac{|\mathcal{H}|+1}{\delta} = o(1)$, which as a reminder means that it has a limit of zero.

Hint: In my solution, $\mathcal{H}$ and δ only appear in the form $\log \frac{|\mathcal{H}|+1}{\delta}$; the 1 isn't some constant hidden by $\mathcal{O}$, it's just a 1.

Hint: Recall that since $\hat{h}_S$ is an ERM, $L_S(\hat{h}_S) \le L_S(h^*)$.

Hint: After doing the things in the hints above, you'll probably get something of the form $L_\mathcal{D}(\hat{h}_S) \le \beta\sqrt{L_\mathcal{D}(\hat{h}_S)} + \gamma$, where β and γ depend on all the other parameters of the problem. Think about what that equation tells us about $L_\mathcal{D}(\hat{h}_S)$, and make your middle school algebra teacher proud.

Answer: TODO

[5] A minimizer is guaranteed to exist, since $\mathcal{H}$ is finite, but at least in my proof it doesn't actually matter that h^* be minimal; you could plug any hypothesis you like into the bound.