CPSC 532D, Fall 2025: Assignment 1
due Monday, 15 September 2025, **11:59 pm**

Do assignment 1 alone; future ones will allow partners. **Read the website section on academic integrity** here for what you're allowed to do and not do; in particular, cite your sources (including people you talked to!) and don't use ChatGPT/etc for anything content-related. If you're not sure if something is okay, ask.

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; feel free to delete whatever boilerplate you want. Or answer in a fresh document if you'd prefer.

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 saves us a surprising amount of grading time).

On the off chance something goes wrong, you can also email your assignment to me directly (`dsuth@cs.ubc.ca`).

1 Loss functions [50 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 $S = (z_1, \ldots, z_m) \sim \mathcal{D}^m$, i.e. the z_i are iid following $\mathcal{D}$.

(1.1) [5 points] Show the empirical risk is an unbiased estimator of the true risk: $\mathbb{E}\, L_S(h) = L_{\mathcal{D}}(h)$ for all $h \in \mathcal{H}$.

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 error rate: models can make 5 guesses at the label, and we count how often the correct label is not any 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 $y_i \in [k]^{h \times w}$ are their corresponding pixel labels.[1] One typical evaluation metric is called mIoU ("mean intersection over union"). One minus the mIOU (to make it a nonnegative "loss" to minimize) is measured on a test set as follows:

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

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 convincing intuitive argument – but a formal proof is one way to be very convincing.)

Answer: TODO

(1.5) [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 find W, we minimize the squared reconstruction error on a training set:

$$\arg\min_{W : WW^\top = I} \sum_{i=1}^{m} \|W^\top W x_i - x_i\|^2. \tag{PCA}$$

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

(1.6) [10 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? You can assume that the maximum-likelihood Gaussian is non-degenerate, i.e. has strictly positive variance.

Answer: TODO

2 Bayes optimality [40 points]

A Bayes-optimal predictor is a predictor which achieves the lowest possible error for *any* function, regardless of a choice of hypothesis class.[2]

We'll consider losses of the form $\ell(h, (x, y)) = l_y(h(x))$, where $h : \mathcal{X} \to \hat{\mathcal{Y}}$ and $l_y : \hat{\mathcal{Y}} \to \mathbb{R}$ for each $y \in \mathcal{Y}$.

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. You might find it helpful to also use $\mathcal{D}_x$ to denote the marginal distribution of x under $\mathcal{D}$.

(2.1) [10 points] Argue that if h and g are predictors such that for every x, $\mathbb{E}_{y\sim\mathcal{F}(x)} l_y(h(x)) \le \mathbb{E}_{y\sim\mathcal{F}(x)} l_y(g(x))$, 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

$$f_{\mathcal{D},l}(x) \in \arg\min_{\hat{y}\in\hat{\mathcal{Y}}} \mathop{\mathbb{E}}_{y\sim\mathcal{F}(x)} l_y(\hat{y}).$$

(2.2) [10 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 $\mathcal{Y} = \{0, 1\}$ under 0-1 loss $l_y(\hat{y}) = \mathbb{1}(y \ne \hat{y})$.

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:

$$l_y(\hat{y}) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 0.001 & \text{if } \hat{y} = 0,\ y = 1 \\ 1 & \text{if } \hat{y} = 1,\ y = 0. \end{cases}$$

Answer: TODO

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

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

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

Answer: TODO

[2] As usual in this course, I'm ignoring issues of measurability; 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.

3 Interpolation learning [10 challenge points]

Assignments in this course will generally have challenge questions. *These questions are harder than the other ones, and worth a total of 10 points, so the effort:points ratio is* much *higher. 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.*

(This one is not that *hard, especially the first parts! Try giving it a shot.)*

Consider a supervised learning setup, where $\mathcal{H}$ contains functions $\mathcal{X} \to \mathcal{Y}$, and the training data is a subset of $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let's restrict ourselves to losses of the form $\ell(h, (x, y)) = l_y(h(x))$ and further assume that $l_y(\hat{y}) \geq 0$ for all $y, \hat{y}$.

Say that a nonnegative loss of this form is *strictly proper* if a sequence of predictions $(\hat{y}_t)$ satisfies $l_y(\hat{y}_t) \to 0$ if and only if $\hat{y}_t \to y$. The squared loss $l_y(\hat{y}) = (y - \hat{y})^2$ is strictly proper; something weird like $l_y(\hat{y}) = \big((y-1) - \hat{y}\big)^2$ is not, and neither is the logistic loss $l_y(\hat{y}) = \log(1 + \exp(-y\hat{y}))$.

(3.1) [2 points] Show that the set of *interpolators*, $G_S = \{h \in \mathcal{H} : L_S(h) = 0\}$, is the same for any strictly proper nonnegative loss.

Answer: TODO

(3.2) [2 points] What does the previous part imply about the set of possible ERMs in $\mathcal{H}$ across multiple strictly proper nonnegative losses?

Answer: TODO

(3.3) [3 points] Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, and $\mathcal{H} = \{h_w = (x \mapsto w \cdot x) : w \in \mathbb{R}^d\}$. Let ℓ be a strictly proper nonnegative loss such that for each $y \in \mathbb{R}$, $l_y : \mathbb{R} \to \mathbb{R}$ is differentiable.

Gather the x_i from S into $\mathbf{X} \in \mathbb{R}^{m \times d}$ and the y_i into $\mathbf{y} \in \mathbb{R}^m$. Suppose that $m < d$ and that $\mathbf{X}$ is of rank m, implying that there are multiple w for which $L_S(h_w) = 0$.

Thus, there are many possible ERM rules; some of them might generalize well and some might not. As we'll explore in more depth later in the course, it's then interesting to ask which solution a particular algorithm will find, so that we can then ask whether we expect that algorithm to work well in general.

Consider the following learning algorithm, a simple version of gradient descent:

- Take as input a starting guess w_0, a learning rate $\eta > 0$, and a dataset $S = ((x_1, y_1), \ldots, (x_m, y_m))$.
- For $t = 0, 1, \ldots$:
  - Let $w_{t+1} = w_t - \eta \frac{1}{m} \sum_{i=1}^m l'_{y_i}(w_t \cdot x_i)\, x_i$.

We'll prove later in class that under some conditions on S, l, and η, this algorithm converges: $w_t \to w_\infty$, where w_∞ is an ERM. Assume that this happens in this situation.

Give a closed-form expression for w_∞ in terms of $\mathbf{X}$, $\mathbf{y}$, and w_0. *This can include things like matrix multiplication and inversion, but no looping,* $\arg\min$*, etc.*

Hint: Consider which directions it is possible for each step to move in, regardless of the particular choice of loss; it will not *be the whole space $\mathbb{R}^d$. This will let you characterize a subset of possible values that the w_t could conceivably take. Reconcile that characterization with the set of w for which $L_S(h_w) = 0$; that will leave you with a matrix expression for w_∞.*

Answer: TODO

(3.4) [3 points] Prove that your answer to the previous part implies $w_\infty = \arg\min_{w : L_S(h_w) = 0} \|w - w_0\|$: we converge to the closest interpolator to where we started.

Hint: This part is fairly "geometric," assuming you got the right answer in the previous part. There are many good approaches, but my preferred one is usually to write things in terms of the singular value decomposition*; there's a brief overview of this, if you're not super familiar, on the course website.*

Answer: TODO