# CPSC 532D, Fall 2022: Assignment 1
## due Tuesday, 20 September 2022, **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.

You can do this with a partner if you'd like (there's a "find a group" post on Piazza). 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.

Submit your answers as a single PDF on Gradescope: link and login instructions on the Canvas site, which you should now be able to get to even if you're not yet officially enrolled yet. Make sure to use the Gradescope group feature if you're working in a group. 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 names 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 Concentrating on concentric circles [40 + 4 bonus points]

In this problem, we'll show that a particular infinite hypothesis class can be PAC-learned with a "direct" proof. *Based in part on SSBD exercise 3.3.*

Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$, and let $\mathcal{H}$ be the class of indicator functions for circles around the origin – that is, $\mathcal{H} = \{h_r : r \in \mathbb{R}_{\geq 0}\}$, where $h_r(x) = \mathbb{1}_{[\|x\| \leq r]}$ (a function which is 1 if $\|x\| \leq r$, 0 otherwise). Use 0-1 loss.

For a given sample $S$, let $r_S = \max_{i:y_i=1} \|x_i\|$, and use $\hat{h}_S$ to denote $h_{r_S}$, the indicator function of a circle with radius $r_S$, the tightest circle containing all of the positive training points.

To start with, let's assume *realizability*: that there is an $h^* \in \mathcal{H}$ such that $L_{\mathcal{D}}(h^*) = 0$.

**(a)** [5 points] Show that $\hat{h}_S$ is an empirical risk minimizer for the hypothesis class $\mathcal{H}$.

Answer: TODO

**(b)** [20 + 4 bonus points] Prove that if we observe $n \geq \frac{1}{\varepsilon} \log \frac{1}{\delta}$ samples from a realizable $\mathcal{D}$, then the probability that $L_{\mathcal{D}}(\hat{h}_S) \geq \varepsilon$ is at most $\delta$.

You may assume that $\mathcal{D}_x$, the distribution of $x$s sampled from $\mathcal{D}$, is continuous. For [4 bonus points], also handle the case where $\mathcal{D}_x$ is not continuous (e.g. if it has a point mass).

*Hint: Three steps: first, what makes a hypothesis have high error in this setting? Next, what would $S$ have to look like in order to get one of those "bad" hypotheses? Last, how likely is it to see an $S$ like that?*

*Hint: A frequently useful inequality is that $1 - a \leq \exp(-a)$.*

*Hint: If you're stuck and want to see something similar-ish (but a little more complicated), check out Example 2.4 of MRT, which is also Exercise 2.3 of SSBD.*

Answer: TODO

Now let's make things a little harder on our learner, by adding random noise. Rather than perfect realizability, let $\mathcal{D}$ be such that $\Pr(y = 1 \mid x) = \begin{cases} 1 - \eta & \text{if } h^*(x) = 1 \\ \eta & \text{if } h^*(x) = 0 \end{cases}$ for some $h^* \in \mathcal{H}$: that is, labels are randomly flipped with probability $\eta \in (0, \frac{1}{2})$. The learner knows the value of $\eta$, but not which points have been flipped.

**(c)** [5 points] Is $\hat{h}_S$ still an ERM?

Answer: TODO

**(d)** [10 points] Ambitious Angus claims to have proven the following:

> For any $\varepsilon, \delta \in (0, 1)$ and $0 \leq \eta < \frac{1}{2}$, there is a function $n_{\mathcal{H}}(\eta, \varepsilon, \delta)$ such that, for any $n \geq n(\eta, \varepsilon, \delta)$ and any $\mathcal{D}$ of the form above,
>
> $$\Pr_{S \sim \mathcal{D}^n}(L_{\mathcal{D}}(\hat{h}_S) > \eta + \varepsilon) \leq \delta.$$

This is followed by an unreadably long computer-assisted proof using both category theory and complicated partial differential equations. Without reading that proof, argue that Angus must be wrong: no such function $n_{\mathcal{H}}$ can exist.

Answer: TODO

## 2 Loss functions [35 points]

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 as we've mostly considered so far are usually 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 $L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, z_i)$ for a sample $S = (z_1, \ldots, z_n) \sim \mathcal{D}^n$. (The notation $\mathcal{D}^n$ here refers to a *product distribution*: a distribution which gives $n$ independent and identically-distributed samples from $\mathcal{D}$.)

**(a)** [10 points] Prove that, for any given $h \in \mathcal{H}$, $L_S$ is unbiased: $\mathbb{E}\, L_S(h) = L_{\mathcal{D}}(h)$.

Answer: TODO

**(b)** [5 points] Prove that the zero-one loss for $k$-way classification ($\mathcal{Y} = \{1, \ldots, k\}$) is equal to one minus the accuracy (the portion of correct answers).

Answer: TODO

**(c)** [5 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

**(d)** [5 points] *Semantic segmentation* is a computer vision problem where we try to label each pixel of an image as belonging to one of several classes ("tree," "street," "dog," etc.). Let $S = ((x_1, y_1), \ldots, (x_n, y_n))$ where $x_i$ are the given input images and $y_i$ their corresponding pixel labels. One typical evaluation metric is called mIoU ("mean intersection over union"), and is measured on a test set as follows:

$$\frac{1}{\#\text{ of classes}} \sum_{k=1}^{\#\text{ of classes}} \frac{\#\text{ of pixels from all images predicted as } k \text{ with label } k}{\#\text{ of pixels from all images predicted as } k \text{ and/or with label } k}.$$

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

**(e)** [5 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:

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

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

**(f)** [5 points] Frame the problem of fitting a Gaussian distribution to a set of scalar observations as loss minimization 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

## 3 Bayes optimality [25 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.[1]

We'll consider loss functions of the form $\ell(h,(x,y)) = \lambda(h(x),y)$, where $h : \mathcal{X} \to \hat{\mathcal{Y}}$ and $\lambda : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$.[2] (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.

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

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

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

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

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

Answer: TODO

**(c)** [5 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:

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

Answer: TODO

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

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

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

Answer: TODO

[1] 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.

[2] This is often how loss functions are defined in the first place; there are a few cases in the course where the more general $\ell$ form is more convenient, but for this question, the $\lambda$ form is a little easier.