CPSC 532S: Assignment 1 – due Thursday, 20 Jan 2022, 11:59pm

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... feel free to 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 not, I'm not printing them out). Or answer in a fresh document; just make it clear which question you're answering where. **If you're submitting to ICML** (or another deadline between January 18th and 31st), you can instead submit handwritten solutions for this assignment only; also note that the lowest assignment will be dropped for your final grade, so you can skip one assignment if you prefer.

You should do this assignment **alone**, and do the whole thing. (The first of these instructions, and probably the second as well, will change for future assignments.) If you look stuff up anywhere other than in SSBD, please **cite your sources**: just say in the answer to that question where you looked. Please do not look at solution manuals / etc for SSBD, or look up proofs of the standard results we're proving in Question 2.

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. You'll be prompted to mark where each question is in your PDF; make sure you mark all relevant pages for each part. (This saves me a surprising amount of time in grading, although it's kind of annoying for this one since Question 2 has a lot of parts; sorry.)

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`).

Some quick meta-notes about the questions (feel free not to read):

Question 1, Some of the Shais' Problems [50 points], lives up to its name; it's a few questions right out of the SSBD book. They, hopefully, should help reinforce the basic definitions / etc. You should be able to do part (a) and probably part (b) now; material for part (c) will be covered in lecture 2.

Question 2, Principal Component Analysis From First Principles [50 points], is about doing some proofs with linear algebra, which will be important later in the course, but mostly doesn't have to do with this first week of the course so much. Only part (a) directly has to do with lecture, and will be covered in class 2; you can do the rest of the question now and come back to that part afterwards. This question is in here (i) to remind you how linear algebra works (since this will matter later), (ii) because I think it's at a nice level of "walk you through a meaningful proof in a way where you do still have to think a bit", and (iii) because it proves some stuff that I think is nice to know that isn't covered in CPSC 340.

1 Some of the Shais' Problems [50 points]

These problems are from the book of Shalev-Shwartz and Ben-David, though I tweaked some notation slightly to agree with what we're using in lecture.

For each of these problems, use the 0-1 loss (the misclassification rate, as discussed in lecture 1 / chapter 2 of SSBD).

(a) [10 points] [SSBD 2.2] Let $\mathcal{H}$ be a class of binary classifiers over a domain $\mathcal{X}$. Let $\mathcal{D}_x$ be an unknown distribution over $\mathcal{X}$, and let f be the target hypothesis in $\mathcal{H}$. Fix some $h \in \mathcal{H}$. Use the 0-1 loss (the misclassification rate). Show that the expected value of $L_S(h)$ over the choice of S equals $L_{\mathcal{D}_x,f}(h)$, namely, $\mathbb{E}[L_S(h)] = L_{\mathcal{D}_x,f}(h)$.

(b) [20 points] [SSBD 3.3] Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0,1\}$, and let $\mathcal{H}$ be the class of concentric circles in the plane – that is, $\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$, where $h_r(x) = \mathbb{1}_{[\|x\| \leq r]}$ (a function which is 1 if $\|x\| \leq r$, 0 otherwise). Prove that $\mathcal{H}$ is PAC learnable (assuming realizability), and its sample complexity is bounded by $n_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{1}{\varepsilon} \log(1/\delta) \right\rceil$.

(c) [20 points] [SSBD 3.7] Show that for every probability distribution $\mathcal{D}$, the Bayes-optimal predictor $f_{\mathcal{D}}$ is optimal, in the sense that for every classifier $g : \mathcal{X} \to \{0,1\}$, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

2 Principal Component Analysis From First Principles [50 points]

We mentioned in class that the loss minimization framework can also apply to cases without a simple label, such as unsupervised learning. Let's use this idea to study perhaps the most common unsupervised linear framework: PCA. We'll introduce it as we go; don't worry if you haven't seen it before.

In PCA, we're given a dataset $X \in \mathbb{R}^{n \times d}$, which contains n points $x_i \in \mathbb{R}^d$ stored as the rows of X.[1]

We're trying to find a linear transformation $W \in \mathbb{R}^{k \times d}$, for $k \leq d$, where W has orthonormal rows ($WW^\mathsf{T} = I_k$), and we want $Wx \in \mathbb{R}^k$ to contain "as much information as it can" about x. For instance, we might want to do this with $k = 2$ to plot high-dimensional data. We can do this to everything in X by taking XW^T; you should make sure it makes sense to you that this is the same as stacking up Wx_i for each i.

To map a point $z \in \mathbb{R}^k$ back to $\mathbb{R}^d$, we just take $W^\mathsf{T} z$ (why?). Thus projecting and reconstructing the entire dataset is $XW^\mathsf{T}W$.

We'll choose W to minimize the squared reconstruction error on our dataset:

$$\arg\min_{W : WW^\mathsf{T} = I_k} \|XW^\mathsf{T}W - X\|_F^2, \tag{PCA}$$

where $\|X\|_F^2 = \sum_{ij} X_{ij}^2 = \operatorname{tr}(X^\mathsf{T}X)$ is the squared Frobenius norm.

You may find it helpful for this question to recall "trace rotation": $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.

(a) [5 points] Frame (PCA) as empirical risk minimization using the terminology from class (also see page 48 of Shai+Shai). Specifically, what are the data domain $\mathcal{Z}$, the sample $S = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, 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)?

Recall that the sample covariance[2] of points $z_1, \ldots, z_n \in \mathbb{R}^d$ is $\frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})(z_i - \bar{z})^\mathsf{T} \in \mathbb{R}^{d \times d}$, where $\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i$ is the sample mean. Another view of PCA is maximizing the variance of the projected points:

(b) [5 points] Suppose, for this part only, that X is centred: $\frac{1}{n} \mathbf{1}_n^\mathsf{T} X = \mathbf{0}_d$, where $\mathbf{1}_n \in \mathbb{R}^n$ is an all-ones vector and $\mathbf{0}_d \in \mathbb{R}^d$ is an all-zeros vector. (This is standard in PCA; it gives us more flexibility in the fit.) Show (PCA) is equivalent to maximizing the trace of the sample covariance of XW^T.

Even requiring orthonormal W, there are many equivalent solutions to (PCA).[3] To reduce the number of valid solutions (and because it has its own advantages), we usually require that W be consistent with PCA using *sequential fitting*: first, we solve (PCA) for $k = 1$. Once we've fit on $j - 1$ components, to get the jth we remove the already-fit component, $X - XW^\mathsf{T}W$, and solve (PCA) with $j = 1$ on that remainder, making that the jth row of W. We'll prove shortly that this doesn't hurt our reconstruction performance.

Thus, for the next little bit, we're going to think only about the solution for $k = 1$. For notational convenience, let the vector $w \in \mathbb{R}^d$ be the first (and only) row of W, so that $W = w^\mathsf{T}$. Plugging in to part (b), we obtain that solutions to (PCA) are exactly the maximizers of $\|Xw\|$. (There's a bit of a hint for part (b)!)

What w are these? We can answer that with the following result, remembering $\|Xw\|^2 = w^\mathsf{T}(X^\mathsf{T}X)w$:

Proposition 2.1. *Let $A \in \mathbb{R}^{m \times m}$ have real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m$, with corresponding real orthonormal eigenvectors q_i. Let k be the largest index such that $\lambda_1 = \lambda_k$ (i.e. the top k eigenvalues are the same). Then*

$$\max_{v : \|v\| = 1} v^\mathsf{T} A v = \lambda_1;$$

[1] As usual in machine learning, we'll think of these as column vectors, even though they're the rows of X. As a student in CPSC 340 said last term, "that's dumb," but oh well.

[2] This is the unbiased sample covariance; the same is true for the biased estimator.

[3] Note that $(RW)^\mathsf{T}(RW) = W^\mathsf{T}(R^\mathsf{T}R)W$ and $(RW)(RW)^\mathsf{T} = RWW^\mathsf{T}R^\mathsf{T} = RR^\mathsf{T}$, so any unitary matrix (where $R^\mathsf{T}R = I = RR^\mathsf{T}$) transforms any solution W into an equivalent one. For instance, you can rotate, permute, or flip signs of the components.

this max is achieved if and only if v is a unit vector in $\text{span}(\{q_1, \ldots, q_k\}) = \{\sum_{i=1}^k \alpha_i q_i : \alpha_i \in \mathbb{R}\}$.

Proof. Pick $\alpha \in \mathbb{R}^m$ such that $v = \sum_{i=1}^m \alpha_i q_i$.

(c) [5 points] Prove this.

□

Recall that the singular value decomposition, the SVD,[4] of a real rank-r matrix $A \in \mathbb{R}^{a \times b}$ is $A = U\Sigma V^\mathsf{T} = \sum_{i=1}^r \sigma_i u_i v_i^\mathsf{T}$, where Σ is a diagonal matrix with entries $\sigma_1 \geq \cdots \geq \sigma_r > 0$, $U \in \mathbb{R}^{a \times r}$ has orthonormal columns $u_i \in \mathbb{R}^a$ so that $U^\mathsf{T}U = I_a$, and $V \in \mathbb{R}^{b \times r}$ has orthonormal columns $v_i \in \mathbb{R}^b$ so that $V^\mathsf{T}V = I_b$. The σ_i are the square roots of the nonzero eigenvalues of $A^\mathsf{T}A$, which coincide with those of AA^T. U contains the corresponding eigenvectors of AA^T, while V has those of $A^\mathsf{T}A$.

Note that the SVD is not unique: you can always flip the signs of both u_i and v_i, and if there are non-distinct singular values, you can rotate the singular vectors in the corresponding subspace.

Let $A_k = \sum_{i=1}^k \sigma_i u_i v_i^\mathsf{T} = U_{1:k}\Sigma_{1:k}V_{1:k}^\mathsf{T}$ be the rank-k approximation obtained by truncating the SVD. Here $U_{1:k}$ and $V_{1:k}$ take the first k columns of U or V.

(d) [10 points] Show that the kth principal component of X, obtained via the sequential algorithm described above, is a valid choice for the kth right singular vector, so $W = V_{1:k}^\mathsf{T}$ for one valid choice of V. Furthermore, show $XW^\mathsf{T}W = X_k$.

We now know that for any k, the PCA reconstructions $XW^\mathsf{T}W$ are equal to X_k. That this is a global minimum for (PCA) is a consequence of the following theorem:

Theorem 2.2 (Eckart-Young, Frobenius)**.** *For any matrix B of rank at most k, $\|A - A_k\|_F \leq \|A - B\|_F$.*

We're now going to prove this theorem together. To do it, we'll first prove the version for the spectral norm (also known as the operator norm). Recall that the spectral norm is $\|A\|_{op} = \sup_x \|Ax\|/\|x\|$, which Proposition 2.1 implies is equal to the largest singular value of A.

Theorem 2.3 (Eckart-Young, operator)**.** *For any matrix B of rank at most k, $\|A - A_k\|_{op} \leq \|A - B\|_{op}$.*

Proof. Recall that A is $a \times b$ and of rank r. Assume, without loss of generality, that $k < r \leq b \leq a$.

(e) [3 points] Explain why this doesn't lose generality.

(f) [2 points] Show that $\|A - A_k\|_{op} = \sigma_{k+1}$.

Hint: You probably already basically did this as part of part (d).

Now, assume for the sake of contradiction that there is some B of rank at most k with $\|A - B\|_{op} < \sigma_{k+1}$.

(g) [5 points] Let y be in the null space of B, i.e. $By = 0$, with $y \neq 0$. Show $\|Ay\| < \sigma_{k+1}\|y\|$.

(h) [5 points] Let $z \in \text{span}(\{v_1, \ldots, v_{k+1}\})$. Show $\|Az\| \geq \sigma_{k+1}\|z\|$.

(i) [5 points] Argue that this is a contradiction.

Hint: Try a dimension-counting argument.

□

Okay – almost done! The theorem we actually wanted to show was:

Theorem 2.2 (Eckart-Young, Frobenius)**.** *For any matrix B of rank at most k, $\|A - A_k\|_F \leq \|A - B\|_F$.*

[4] Here we're using the "compact" SVD, which I just find a little more convenient to think about. Everything would be basically the same with the "full" SVD.

Proof. Let B be any $a \times b$ matrix of rank at most k.

(j) [5 points] Prove the theorem.

Hint: Try using Theorem 2.3, recalling that

$$\|A\|_F^2 = \mathrm{tr}(A^\mathsf{T}A) = \mathrm{tr}(V\Sigma U^\mathsf{T}U\Sigma V^\mathsf{T}) = \mathrm{tr}(\underbrace{V^\mathsf{T}V}_{I}\Sigma\underbrace{U^\mathsf{T}U}_{I}\Sigma) = \mathrm{tr}(\Sigma^2) = \sum_{i=1}^{r}\sigma_i^2.$$

□

Alternatively, if you want – it's basically the same proof – you could show instead:

Theorem 2.4 (Eckart-Young, more general)**.** *Let $f(A) = f(\sigma_1, \ldots, \sigma_{\max(a,b)})$ be a function of the singular values of A (where $\sigma_i = 0$ if $i > r$). Suppose that if $\sigma'_i \geq \sigma_i$ for all i, then $f(\sigma'_1, \ldots, \sigma'_{\max(a,b)}) \geq f(\sigma_1, \ldots, \sigma_{\max(a,b)})$. Then, for all matrices B of rank at most k, $f(A - A_k) \leq f(A - B)$.*

In particular, this will include all *unitarily invariant* matrix norms, i.e. matrix norms such that $\|RAS\| = \|A\|$ for all unitary matrices R and S.