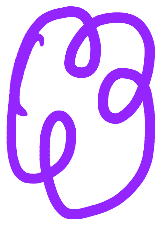
CS 446 / ECE 449: Machine learning.
Exam #1: due at noon, CST, on March 18, 2021.

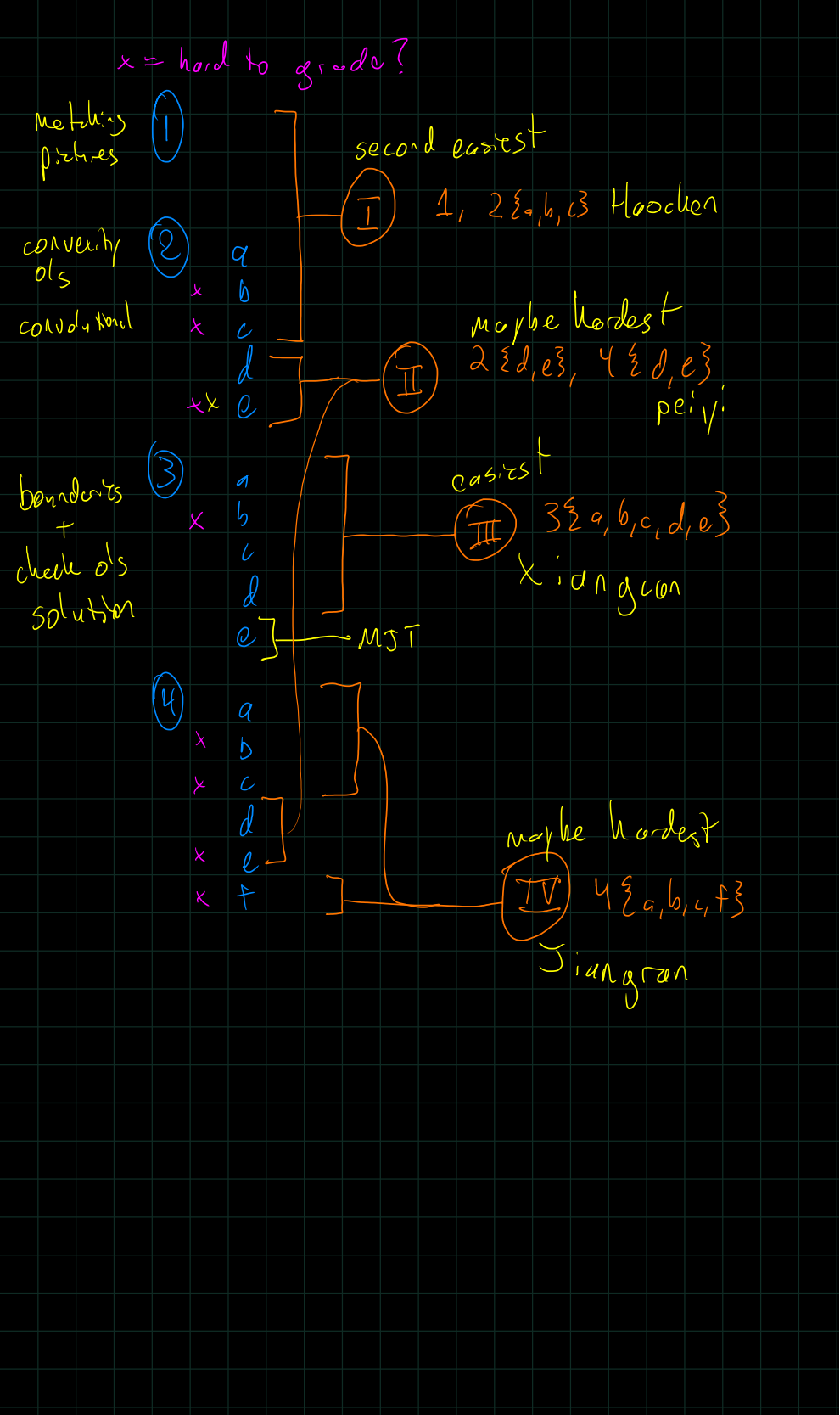


Name:	MJT
NetID:	

- **This is an individual exam:** you may not consult with anyone except course staff.
- **Resources you may access:** course lecture slides (`Lxy_Slides_Split.pdf` recommended for fewest typos), your homework solutions, and two single-sided pages typed or hand-written at a reasonable size.
- **Citing lecture:** if you skip steps via something from lecture, be explicit: cite the lecture and slide number.
- **Where to go for clarifications** *(not "hints" or "help"!)*:
  - Make a private (TA/instructor only) post on campuswire.
  - Attend one of the following campuswire office hours (includes extras over regular schedule): **3/16:** 15:30 – 17:30, 22:00 – 22:30; **3/17:** 11:00 – 11:30, 16:30 – 17:30; **3/18:** 10:00 – 11:00.
- **Where *not* to go for help** includes (but is not limited to):
  - Discord.
  - Public posts on campuswire.
  - Any discussion with other students.
  - ...
- **How to submit:**
  1. Prepare a PDF of your solutions; e.g., (a) print the exam, write on it, and take pictures, (b) write directly on the exam via a tablet, (c) write on blank sheets and take pictures.
  2. Upload to gradescope under `midterm`, similarly to homeworks. You may re-submit often.
- **Double-ruled boxes:** if you write directly on the exam PDF, please limit your answers to these boxes. They were designed to give generous amounts of space.
- **Point values:** problems have different point values (as indicated), but parts of problems are all equal! Do not lose too much time on parts that are much more difficult!
- **Notation:** $\|\cdot\|$ without a subscript means 2-norm (Euclidean norm), thus $\|\boldsymbol{v}\|^2 = \|\boldsymbol{v}\|_2^2 = \sum_{i=1}^d v_i^2$. "$\boldsymbol{x}$ has unit norm" means $\|\boldsymbol{x}\| = 1$. $\boldsymbol{e}_i$ is the i^{th} standard basis vector of appropriate dimension. k-nearest-neighbor uses 2-norm by default. The decision boundary of a linear SVM with no bias goes through the origin; with a bias, it need not pass through the origin.

---

x = hard to grade?

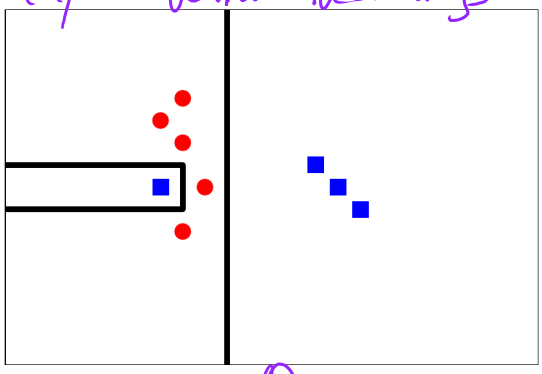


1. Decision boundaries (6 total points, 1 point each figure).

Match the following predictors to the following decision boundaries: (a) 1-nearest-neighbor, (b) axis-aligned decision tree, (c) linear SVM, (d) arbitrary linear separator, (e) RBF SVM, (f) ReLU network. Note that matching means you use each answer exactly once; there is only one valid matching.

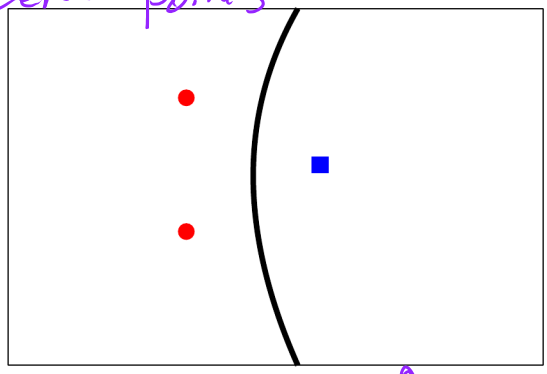
only valid matchings receive points

1.1



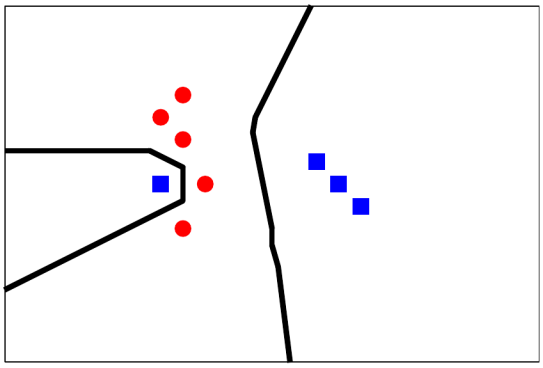
Circle one: (a) (b) (c) (d) (e) (f)

1.4



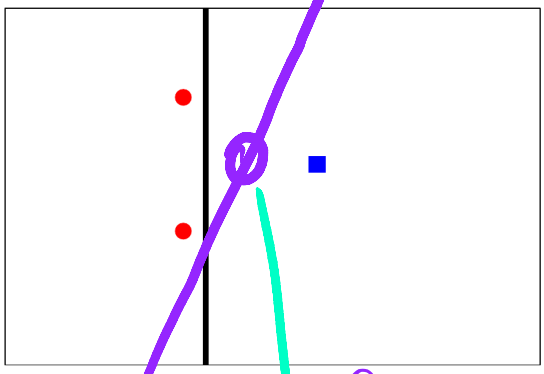
Circle one: (a) (b) (c) (d) (e) (f)

1.2



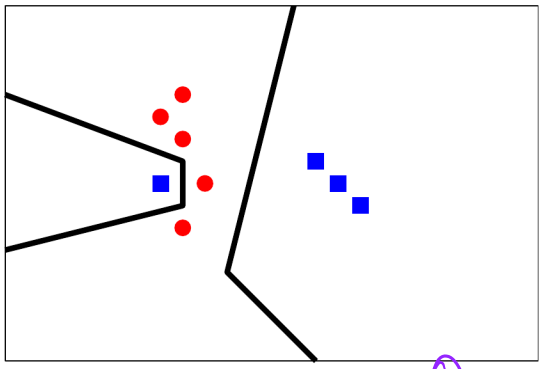
Circle one: (a) (b) (c) (d) (e) (f)

1.5



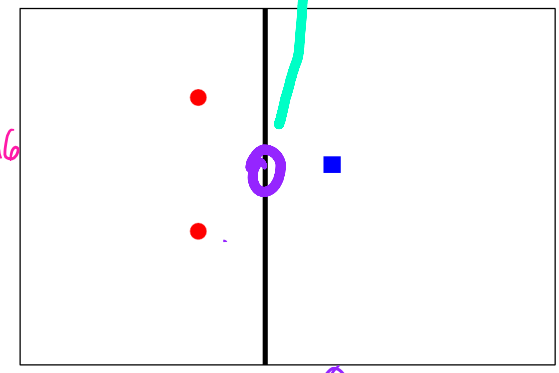
Circle one: (a) (b) (c) (d) (e) (f)

1.3



Circle one: (a) (b) (c) (d) (e) (f)

1.6



Circle one: (a) (b) (c) (d) (e) (f)

partial credit delicate

2. Short answer (15 total points, 3 per part).

(a) Prove that $g(x) := x^3$ is not convex (over $\mathbb{R}$).

Hint: Produce a counterexample to the definition on slide 5/26 from lecture 5. (Your proof needs to be an explicit symbolic derivation, but consider first studying a plot on scratch paper.)

Correct counterexample but no proof: total = 1 points.

Consider $x := -1$, $x' := 0$, $\alpha := \frac{1}{2}$,

and $z := \alpha x + (1-\alpha) x' = -\frac{1}{2}$.

Then $g(z) = -\frac{1}{8} > -\frac{1}{2} = \alpha(-1) + (1-\alpha)(0)$

$= \alpha g(x) + (1-\alpha) g(x')$,

meaning g not convex.

Can also do other points, also g' & g''.

(b) Consider the following least squares problem: data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ has rank $r < \min\{n, d\}$, target vector $\boldsymbol{y} \in \mathbb{R}^n$ is given, and let $\hat{\boldsymbol{w}}_{\text{ols}}$ denote the OLS solution. Let $\boldsymbol{v}$ denote any nonzero vector which is orthogonal to the right singular vectors $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r)$ of $\boldsymbol{X}$, and define another solution $\boldsymbol{w} := \hat{\boldsymbol{w}}_{\text{ols}} + \boldsymbol{v}$. Show that $\boldsymbol{w}$ is an optimal solution to this least squares problem.

As in lecture 2, it suffices to show that w satisfies the normal equations.

Let $X = \sum_{i=1}^{r} s_i u_i v_i^T$ be an SVD with the provided right singular vectors, whereby $Xv = \sum_{i=1}^{r} s_i u_i v_i^T v = 0$. Since $\hat{w} := \hat{w}_{ols}$ satisfies the normal equations,

$X^T y = X^T X \hat{w} = X^T (X\hat{w} + 0) = X^T (X\hat{w} + Xv) = X^T X (\hat{w} + v) = X^T X w.$

seems there's a risk of cheating in here?

(c) Consider a convolutional layer with one 2x2 filter $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$, all entries nonzero, and the simplest settings: no bias, no padding, a stride of 1, no dilation. Construct a 3x3 input image, not all entries equal to zero, which is in the *null space* of this filter, meaning the output image after applying the above filter is all zeros. For full points, include your input image and a brief verification that the output image is all zeros.

Either convolution convention okay (filter reversed or not).

Consider input data $X = \begin{bmatrix} c & 0 & 0 \\ -a & 0 & 0 \\ a^2/c & 0 & 0 \end{bmatrix}$.

Then

$$\text{conv}\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}, X\right) = \left[\begin{array}{c|c} ac-ac & 0 \\ \hline -a^2+c\left(\frac{a^2}{c}\right) & 0 \end{array}\right] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

If fixed filter, then ≤ 1 total point

tough to grade. Many crazy solutions.

(d) Consider the *three-dimensional* XOR problem in $\mathbb{R}^3$: there are 8 data points, consisting of all possible $(\pm 1, \pm 1, \pm 1)$ triples, and the labels are the products of the three coordinates. Circle all of the classifiers/methods in the following list which can perfectly classify this problem:

i. 2-layer ReLU network,	**iv.** decision tree with 3 axis-aligned splits,
ii. fourth degree polynomial kernel SVM,	**v.** 1-nearest-neighbor,
iii. linear predictor,	**vi.** 4-nearest-neighbor,

-1 point: 1-2 errors

-2 points: 3-4 errors

-3 points 5-6 4 errors.

(e) Recall the primal and dual optimization problems corresponding to the hard-margin SVM for linearly separable data:

$$\min\left\{\frac{1}{2}\|\boldsymbol{w}\|^2 : \boldsymbol{w}\in\mathbb{R}^d, \forall i \,.\, \boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i y_i \geq 1\right\} = \max_{\boldsymbol{\alpha}\in[0,\infty)^n} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j \boldsymbol{x}_i^\mathsf{T}\boldsymbol{x}_j.$$

Consider the following (linearly separable) data: $d=2$, $n=2$, $\boldsymbol{x}_1=\boldsymbol{e}_1$, $y_1=+1$, $\boldsymbol{x}_2=-\boldsymbol{x}_1=-\boldsymbol{e}_1$, $y_2=-y_1=-1$. Show that the dual optimum is not unique.

Hint: Here is one way to solve the problem (there are others). On scratch paper, draw the data and guess an optimal primal weight vector via the geometric view of the SVM (this primal weight vector will be unique). Then write out the dual, plug in the data, and solve for two distinct optimal values of $\boldsymbol{\alpha}$. To write up your proof for your final solution, you can produce your primal solution and two distinct dual solutions, point out that the corresponding primal and dual objective values are equal (thus all three values are optimal), which implies the dual optimum is not unique.

Hard part is showing distinct dual choices are optimal, here I use weak duality!

Or can check with gradients equal 0.

Define $\bar{w} := e_1$, } primal.

$\bar{\alpha} := e_1$ } dual.

$\bar{\beta} := e_2$ }

$\bar{w}$ is feasible: $\bar{w}^T e_1 (+1) \geq 1$, $\bar{w}^T (-e_1)(-1) \geq 1$.

primal value: $\frac{1}{2}\|\bar{w}\|^2 = \frac{1}{2}$.

Dual values: $D(\bar{\alpha}) = 1 - \frac{1}{2} y_1^2 x_1^T x_1 = \frac{1}{2}$,

$D(\bar{\beta}) = 1 - \frac{1}{2} y_2^2 x_2^T x_2 = \frac{1}{2}$.

$\frac{1}{2}\|\bar{w}\|^2 = D(\bar{\alpha}) = D(\bar{\beta})$, thus respectively primal, dual, dual by weak duality → optimal, and note $\bar{\alpha} \neq \bar{\beta}$.

Many little details. Any other approach quite messy & error prone. Optimality conditions delicate since primal & dual are constrained.

two ways to solve

① [via hint.]

Start from distinct feasible dual variables $\vec{\alpha} \neq \vec{\beta}$,

then argue they are optimal.

maybe take off points

In reference solution above, I used $\vec{w}$ & weak duality to assert their optimality. //

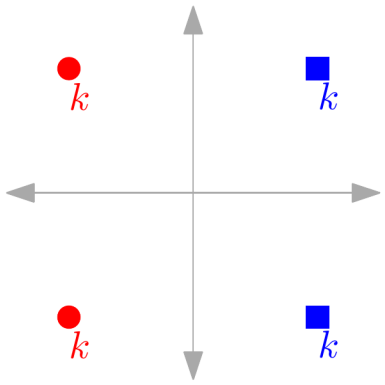
② Take gradient of dual objective, set to zero, give two different solutions

(I'm fine with this approach.)

3. Robustness (21 total points, 3 per part).

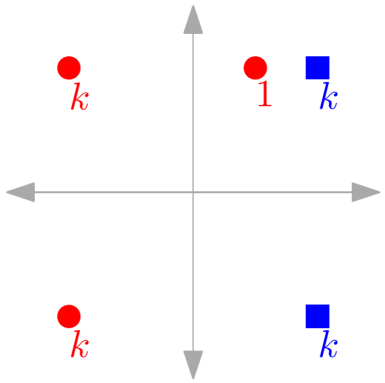
In this problem, we will study how different predictors respond to slight changes in the training data. We will use two sets of training data, one with $4k$ points and another with $4k+1$ points, where $k \geq 3$ is arbitrary.

- **Training set 1, used in parts (a) & (b).** This data consists of $n_1 := 4k$ points in $\mathbb{R}^2$, forming the rows of input matrix $\boldsymbol{X}_1 \in \mathbb{R}^{n_1 \times 2}$, with corresponding label vector $\boldsymbol{y}_1 \in \{\pm 1\}^{n_1}$.



  In detail, this data has $k \geq 3$ copies of each of the four points of the form $(\pm 2, \pm 2)$. The labels are given by the colors/shapes: let red/circle denote -1, and blue/square denote $+1$. This data is linearly separable.

- **Training set 2, used in parts (c) – (g).** This data consists of $n_2 := 4k+1$ points in $\mathbb{R}^2$, forming the rows of input matrix $\boldsymbol{X}_2 \in \mathbb{R}^{n_2 \times 2}$, with corresponding label vector $\boldsymbol{y}_2 \in \{\pm 1\}^{n_2}$.

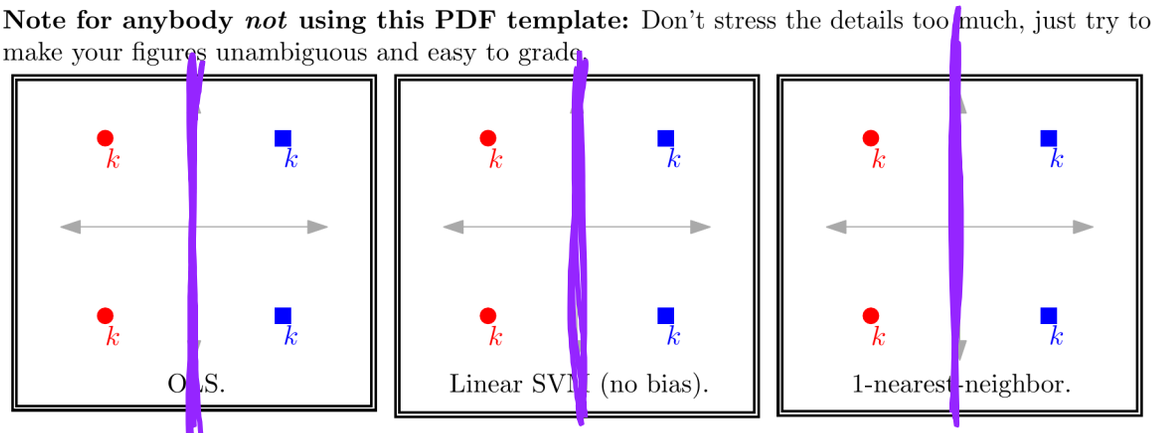


  In detail, $4k$ of the points are common with training set 1, however there is an additional red/circle point (with label -1) at location $(1, 2)$. The data is still linearly separable.

These parts of the problem use training set 1, meaning inputs $\boldsymbol{X}_1$ and labels $\boldsymbol{y}_1$.

(a) *Draw* the decision boundary of OLS, linear SVM (with no bias), and 1-nearest-neighbor in the following three plots. You do not need to justify your answers.

Note for anybody *not* using this PDF template: Don't stress the details too much, just try to make your figures unambiguous and easy to grade.



(b) Write down the OLS solution,and provide a verification that it satisfies the normal equations.

Hint: You don't need to do tedious calculations for this problem, here are two ways to do it. The first option is to compute all objects in the normal equations, and from there compute a solution and argue it is actually the OLS solution. A second option is to "guess" the SVD of $\boldsymbol{X}_1$, which has a clean form, and from there get the OLS solution.

Wlog $X = \begin{bmatrix} 2 & 2 \\ 1 & 1 \\ 2 & -2 \\ 1 & -1 \\ -2 & 2 \\ -1 & 1 \\ -2 & -2 \\ -1 & -1 \end{bmatrix}$ $\begin{matrix} \} 4 \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix}$, $y = \begin{bmatrix} +1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$ $\begin{matrix} \} 2k \\ \\ \} 2k. \\ \\ \end{matrix}$

X has orthogonal columns, thus

$X^T X = \begin{bmatrix} 16k & 0 \\ 0 & 16k \end{bmatrix}$, which is invertible,

therefore normal equations have unique solution $(X^T X)^{-1}(X^T)y = \begin{bmatrix} 1/16k & 0 \\ 0 & 1/16k \end{bmatrix}\begin{bmatrix} 8k \\ 0 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 0 \end{bmatrix}$, which must agree with ols solution.

(see next page for another.)

↳ from lecture: ols always satisfies normal.

Note $X = s_1 u_1 v_1^T + s_2 u_2 v_2^T$ where

$$v_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix} \qquad v_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}$$

$$u_1 = \frac{1}{\sqrt{2k}}\begin{bmatrix}1\\1\\0\\\vdots\\0\\-1\\-1\end{bmatrix}\Big\}k, \qquad u_2 = \frac{1}{\sqrt{2k}}\begin{bmatrix}0\\1\\1\\-1\\-1\\0\end{bmatrix}\Big\}k$$

$$s_1 = 4\sqrt{k}, \qquad s_2 = 4\sqrt{k}.$$

Note also

$$u_1^T y = \sqrt{2k}, \qquad u_2^T y = \sqrt{2k},$$

and

$$\hat{w}_{ols} := X^+ y = \frac{1}{s_1} v_1 u_1^T y + \frac{1}{s_2} v_2 u_2^T y$$

$$= \frac{1}{4\sqrt{k}} \cdot \frac{1}{\sqrt{2}}\left(\begin{bmatrix}1\\1\end{bmatrix}\sqrt{2k} + \begin{bmatrix}1\\-1\end{bmatrix}\sqrt{2k}\right) = \frac{1}{4\sqrt{2k}}\begin{bmatrix}2\sqrt{2k}\\0\end{bmatrix}$$

$$= \begin{bmatrix}1/2\\0\end{bmatrix}. \;//$$

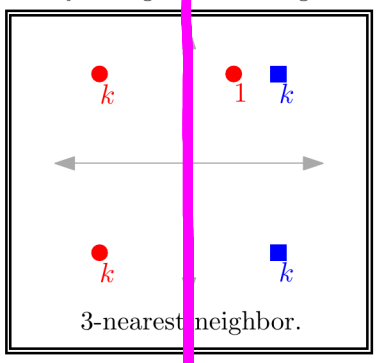
(Quite annoying!)

also, must say "was proved in lecture that ols satisfies normal equations", or check them.

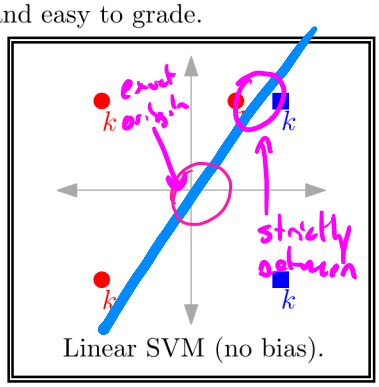
These parts of the problem use training set 2, meaning inputs X_2 and labels y_2.

(c) *Draw* the decision boundary of 3-nearest-neighbor, linear SVM (with no bias term), and 1-nearest-neighbor in the following three plots. You do not need to justify your answer.

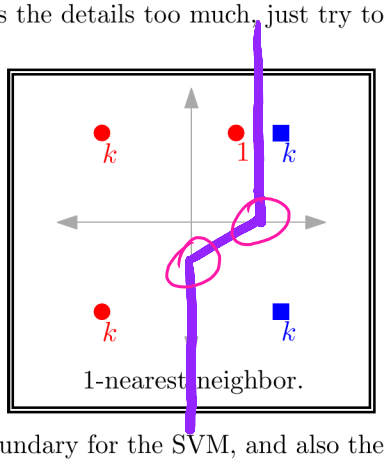
Note for anybody *not* using this PDF template: Don't stress the details too much, just try to make your figures unambiguous and easy to grade.



3-nearest neighbor.



Linear SVM (no bias).



1-nearest neighbor.

(d) State the number of points within a distance of 1 of the decision boundary for the SVM, and also the number of points within a distance of 1 of the decision boundary for the 3-nearest-neighbor classifier.

SVM: 2k+1 points.

3-nn: 1 point.

(e) Limiting yourself to 1–5 sentences, state in intuitive terms whether you prefer the linear SVM solution or the 3-nearest-neighbor solution.

This problem will be graded leniently, but for full points, refer to parts (e) and (f), and how you feel they may relate to the performance on (unseen!) testing data.

If data (seen & unseen) has small perturbations & label flips, SVM is unstable (as above) & 3-nn is ideal (low test error).

If data dist non-noisy, linear SVM could be good in some low-sample regimes.

4. Vanishing/exploding gradients (18 total points, 3 per part).

Consider an L-layer ReLU network with input dimension, output dimension, and all widths equal to 2, and no biases, meaning it computes $f(\boldsymbol{x}; \mathcal{W})$ given by

$$\mathbb{R}^2 \ni \boldsymbol{x} \mapsto f(\boldsymbol{x}; \mathcal{W}) := \boldsymbol{W}_L \sigma(\boldsymbol{W}_{L-1} \cdots \sigma(\boldsymbol{W}_1 \boldsymbol{x}) \cdots) \in \mathbb{R}^2,$$

where each $\boldsymbol{W}_i \in \mathbb{R}^{2\times 2}$, and each $\sigma : \mathbb{R}^2 \to \mathbb{R}^2$ applies a ReLU coordinate-wise (meaning $\sigma(\boldsymbol{v}) = (\max\{0, v_1\}, \max\{0, v_2\})^\intercal \in \mathbb{R}^2$), and $\mathcal{W} = (\boldsymbol{W}_L, \ldots, \boldsymbol{W}_1)$ is notation for the tuple of parameters across layers.

Throughout this problem, you will be asked to choose parameters $\mathcal{W} = (\boldsymbol{W}_L, \ldots, \boldsymbol{W}_1)$ which satisfy certain conditions. In all parts of this problem:

i. An input vector $\boldsymbol{x} \in \mathbb{R}^2$ is given with unit norm, but is otherwise arbitrary;

ii. You must choose matrices $\mathcal{W} = (\boldsymbol{W}_L, \ldots, \boldsymbol{W}_1)$, where each matrix $\boldsymbol{W}_i \in \mathbb{R}^{2\times 2}$ has rank 2, with two distinct singular values exactly equal to 2 and 1/2 respectively;

iii. You must provide both an answer and a brief derivation/verification for full credit.

The parts of the problem are as follows.

(a) Let unit norm input $\boldsymbol{x}$ be given, and let $\boldsymbol{z} \in \mathbb{R}^2$ be another unit norm vector which is orthogonal to $\boldsymbol{x}$. Define a matrix

$$\mathbb{R}^{2\times 2} \ni \boldsymbol{M} := 2\boldsymbol{e}_1 \boldsymbol{x}^\intercal + \frac{1}{2}\boldsymbol{e}_2 \boldsymbol{z}^\intercal.$$

Provide the singular values of $\boldsymbol{M}$, and compute $\boldsymbol{M}\boldsymbol{x}$ and $\boldsymbol{M}\boldsymbol{z}$.

Hint: This is not a trick question; compare the definition of $\boldsymbol{M}$ to the definition of the SVD from slide 9/26 in lecture 2.

e_1 & e_2 are orthonormal, as are x & z, therefore $2e_1x^\intercal + \frac{1}{2}e_2z^\intercal$ is an SVD for M, and singular values are 2 & $\frac{1}{2}$.

Moreover

$$Mx = 2e_1x^\intercal x + \frac{1}{2}e_2z^\intercal x = 2+0 = 2,$$

$$Mz = 2e_1x^\intercal z + \frac{1}{2}e_2z^\intercal z = 0+\frac{1}{2} = \frac{1}{2}.$$

(b) Let unit norm input $\boldsymbol{x}$ be given. Provide a choice of $\mathcal{W}$ so that $\|f(\boldsymbol{x};\mathcal{W})\| = 2^L$.
Hint: The previous part can help you choose $\boldsymbol{W}_1$ (though, if you use it, you must still provide verification); how should you choose the weight matrices in other layers?

Choose $W_1 = M = 2e_1 x^T + \frac{1}{2} e_2 z^T$

where z unit & $x^T z = 0$.

Choose $W_i := 2e_1 e_1^T + \frac{1}{2} e_2 e_2^T$, $i \geq 2$.

By construction, singular values $(2, \frac{1}{2})$.

Moreover $W_L \sigma(W_{L-1} \cdots W_3 \sigma(W_2 \sigma(W_1 x)) \cdots) = \begin{bmatrix} 2^L \\ 0 \end{bmatrix}$,

which has norm 2^L ... $\in \begin{bmatrix} 2 \\ 0 \end{bmatrix} \in \begin{bmatrix} 4 \\ 0 \end{bmatrix} \in \begin{bmatrix} 8 \\ 0 \end{bmatrix}$

no need for induction

at most 2 point if no verification

(c) Let unit norm input $\boldsymbol{x}$ be given. Provide a choice of $\mathcal{W}$ so that $\|f(\boldsymbol{x};\mathcal{W})\| = 2^{-L}$.

Pick z as before, define

$W_1 := 2e_1 z^T + \frac{1}{2} e_2 x^T$

and $W_i := 2e_1 e_1^T + \frac{1}{2} e_2 e_2^T$,

satisfying conditions by construction.

By induction, output of W_i is $\begin{bmatrix} 0 \\ 2^{-i} \end{bmatrix}$:

Base: $W_1 x = 0 + \frac{1}{2} e_2$; IS: $W_i \sigma\left(\begin{bmatrix} 0 \\ 2^{-i} \end{bmatrix}\right) = W_i 2^{-i} e_2$

norm is 2^{-i}. $= 2^{-i} e_2$

(d) Let unit norm input $\boldsymbol{x}$ be given, and let $\boldsymbol{W}_1 \in \mathbb{R}^{2\times 2}$ be any 2×2 matrix satisfying the conditions on the first layer which hold throughout this problem. Show that it is impossible to have $\boldsymbol{W}_1\boldsymbol{x} = 0$.
Hint: What is the rank of $\boldsymbol{W}_1$? What is the largest possible rank of a 2×2 matrix?

lenient

any proof / argument is fine

W_1 is full rank, thus has no non-trivial right null space.

OR:

Implies $x = W^{-1}Wx = W^{-1}0 = 0$,

but $\|x\| = 1$, contradiction.

(e) Let unit norm input $\boldsymbol{x}$ be given. Provide a choice of $\mathcal{W}$ so that $\|f(\boldsymbol{x}; \mathcal{W})\| = 0$.
Hint: You can't enforce $\boldsymbol{W}_1\boldsymbol{x} = 0$, since that would contradict the previous part. Keep in mind what you have extra in this part: you have L layers, not just one, and you have ReLUs.

} $L > 1$

No need for explicit induction

Choose z as ~~before~~ and $W_1 := -2e_1x^T + \frac{1}{2}e_2z_1^T$

and $W_i := 2e_1e_1^T + \frac{1}{2}e_2e_2^T$ as before;

all conditions met, and

$$\sigma(W_1x) = \sigma\left(\begin{bmatrix}-2\\0\end{bmatrix}\right) = \begin{bmatrix}0\\0\end{bmatrix}, \text{ and}$$

$$\|W_L\sigma(\cdots \underbrace{W_2\sigma(W_1x)}_{\begin{bmatrix}0\\0\end{bmatrix}}\cdots)\| = 0.$$

(f) In lecture we only wrote out an abstract form of a deep network gradient; here is a clean explicit way for the simple networks in this problem. First, define *diagonal activation matrices* $(\boldsymbol{D}_L, \ldots, \boldsymbol{D}_1)$ as

$$\begin{aligned}
\boldsymbol{D}_1 &:= \operatorname{diag}(\mathbb{1}[\boldsymbol{W}_1 \boldsymbol{x} \geq 0]) \in \mathbb{R}^{2\times 2}, \\
\boldsymbol{D}_2 &:= \operatorname{diag}(\mathbb{1}[\boldsymbol{W}_2 \boldsymbol{D}_1 \boldsymbol{W}_1 \boldsymbol{x} \geq 0]) \in \mathbb{R}^{2\times 2}, \\
&\vdots \\
\boldsymbol{D}_i &:= \operatorname{diag}(\mathbb{1}[\boldsymbol{W}_i \boldsymbol{D}_{i-1} \cdots \boldsymbol{D}_1 \boldsymbol{W}_1 \boldsymbol{x} \geq 0]) \in \mathbb{R}^{2\times 2}, \\
&\vdots \\
\boldsymbol{D}_{L-1} &:= \operatorname{diag}(\mathbb{1}[\boldsymbol{W}_{L-1} \boldsymbol{D}_{L-2} \cdots \boldsymbol{D}_1 \boldsymbol{W}_1 \boldsymbol{x} \geq 0]) \in \mathbb{R}^{2\times 2}.
\end{aligned}$$

(These matrices depend on both $\boldsymbol{x}$ and $\mathcal{W}$, but we hide this notation.) The gradient with respect to layer i of the first output coordinate of the whole network, can be written as

$$\frac{\mathrm{d} f(\boldsymbol{x}; \mathcal{W})_1}{\mathrm{d} \boldsymbol{W}_i} = \left(\boldsymbol{e}_1^\mathsf{T} \boldsymbol{W}_L \boldsymbol{D}_{L-1} \cdots \boldsymbol{W}_{i+1} \boldsymbol{D}_{i+1}\right)^\mathsf{T} \left(\boldsymbol{D}_{i-1} \boldsymbol{W}_{i-1} \cdots \boldsymbol{D}_1 \boldsymbol{W}_1 \boldsymbol{x}\right)^\mathsf{T}.$$

(You don't need to verify this formula in this problem, you just need to apply it.)

Let unit norm input $\boldsymbol{x}$ be given, suppose $L > 2$, and consider $\frac{\mathrm{d} f(\boldsymbol{x}; \mathcal{W})_1}{\mathrm{d} \boldsymbol{W}_2}$, the gradient with respect to $\boldsymbol{W}_2$ (of the first output coordinate), the parameters in layer 2, as given by the earlier formula. Provide a choice of $\mathcal{W}$ so that

$$\left\| \frac{\mathrm{d} f(\boldsymbol{x}; \mathcal{W})_1}{\mathrm{d} \boldsymbol{W}_2} \right\|_F = 2^{L-1}.$$

Hint: Try your solution from part (b).

No verification give at most 1 point.

Choose solution in part (b)

By induction, $z_{i-1} = \sigma(W_{i-1} z_{i-2}) = \begin{bmatrix} 2^{i-1} \\ 0 \end{bmatrix}$,

$$D_{i-1} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Thus

$$\left\| e_1^T W_L \cdots D_{i+1} \right)^T \left(D_{i-1} W_{i-1} \cdots D_1 W_1 x \right)^T \right\|$$

$$\begin{bmatrix} 2^{L-i} \\ 0 \end{bmatrix}^T \qquad \begin{bmatrix} 2^{i-1} \\ 0 \end{bmatrix} \to 2^{L-1}.$$