CSE 512: Midterm Oct. 18, 2021

Instructions:

- You have 80 continuous minutes to complete the exam, self-timed.
- You are allowed 1 page (front and back) cheat sheet. The cheat sheet must be scanned (or photographed with high resolution) and submitted along with the exam solutions. The time begins when you flip this page.
- You are also allowed a simple calculator. (You may use MATLAB or Python if you agree to only use the simple calculator functions.)
- You may print the exam and write your solutions or use lyx/latex. If you need extra sheets of paper, please label them carefully as to which question they are answering. Make your final answer clear.
- If you choose to handwrite your solutions, you must make sure that the digital scan / photograph is of high enough quality that we can see everything clearly. Anything we can't read, we will not grade.
- You may not discuss any problem with any other student while the exam submission portal is still open. You may not look for answers on the internet or in any notes outside of your cheatsheet.

Name: _____

Student ID: _____

Scoring	
Q 1	_____ / 10
Q 2	_____ / 30
Q 3	_____ / 30
Q 4	_____ / 30
Total	_____ / 100

1. **Concepts.** You do not need to provide a justification for full credit, but we will read it for partial credit.

   (a) **(2 pts) True** or **False**. In the asymptotic regime (number of training samples $\to +\infty$) the 1-NN classifier is always 100% accurate.

   Ans. False. Just because two data samples are identical in features does not necessarily mean they have the same label.

   (b) **(2 pts) True** or **False**. Stochastic gradient descent minimizes convex functions, but gradient descent can only minimize concave functions.

   Ans. False. Either method can be used to minimize convex or concave functions, but in both cases, if a function is not convex then we cannot be certain we will reach a global minimum.

   (c) **(2 pts) True** or **False**. A Bayesian statistical model accounts for model parameter randomness by imposing a prior distribution.

   Ans. True. This is the definition of the Bayesian framework.

   (d) **(2 pts) True** or **False**. Maximum likelihood estimators of statistical model parameters are always unbiased.

   Ans. False. As examples, maximum likelihood estimators for the variance of a Gaussian or exponential distribution are biased.

   (e) **(2 pts) True** or **False**. Two Gaussian random variables are independent if and only if their covariance matrix is diagonal.

   Ans. True.

2. **Decision theory.** I have gone camping, and am now very far form home.

(a) **Make a decision.** I need to decide if I should even try to walk home or call a helicopter to just pick me up. I know that if I am less than 50 miles from home, I could walk home and be tired, but in one piece. If I am more than 50 miles from home, I'll probably die of thirst and starvation before making it home.

The cost of walking home is 0, but the cost of death is $\$10,000$. The cost of calling for a helicopter is \$800, regardless of where I am. Based on my current information, there is a 1% chance that I am more than 50 miles from home.

i. **(5 pts)** Compute the maximum risk of each decision (call a helicopter or walk home). Using a minimax estimator, decide which action I should take.

Ans. Maximum risk of

- Walking home: dying costs \$10,000
- Calling a helicoptor: \$800.

Therefore, a minimax estimator would have me call for a helicopter, which has the minimum max risk.

ii. **(10 pts)** Compute the Bayes risk of each decision. Using a Bayes classifier, decide which action I should take.

Ans. Bayes risk of

- Walking home: $\$10{,}000 \times 1\% + 0 \times 99\% = \$100$.
- Calling a helicoptor: $\$800 \times 1\% + \$800 \times 99\% = \$800$.

A Bayes estimator would then say the chances of me dying are slim enough that I should just tie my shoes and start walking.

(b) In the desert, there are random sensors scattered around, of which know their location, and also take 3 temperature measurements over 3 days. I have access to some of their measurements:

Sensor	Day 1 Temp	Day 2 Temp	Day 3 Temp	distance to home
A	25° C	25° C	27° C	39 miles
B	5° C	−5° C	2° C	9 miles
C	−50° C	−50° C	−50° C	60 miles
D	15° C	16° C	17° C	42 miles
E	22° C	25° C	30° C	90 miles

I also have a thermometer, and each day I have measured the temperature of where I am lost, recorded as

Day 1 Temp	Day 2 Temp	Day 3 Temp
25° C	24° C	23° C

i. **(5 pts)** Calculate the *mean squared error* between each of the sensor measurements and mine; that is,

$$\textbf{MSE}(\text{sensor } S) = \frac{1}{3}\sum_{i=1}^{3}(T_{S,[i]} - \hat{T}_{[i]})^2$$

where

$$T_{S,[i]} = \text{temperature of sensor } S \text{ on day } i, \qquad \hat{T}_{[i]} = \text{my temperature on day } i.$$

Report your answer to the nearest 0.1.

Ans.

sensor	A	B	C	D	E
dist	5.7	560.7	5476.7	66.7	19.7

ii. **(5 pts)** Using your calculations, decide on my **distance** from home, using a KNN classifier, for $K = 1$ and $K = 3$. Use an averaging rule.

Ans. The closest neighbor is A and the closest three are A, D, and E. Therefore a 1-NN classifier gives a distance of 33 miles, and a 3-NN, using an averaging rule, gives

$$(39 + 42 + 90)/3 = 57 \text{ miles}$$

estimate.

iii. **(5 pts)** Using your calculations, **decide if I am more or less than 50 miles from home**, using a KNN classifier, for $K = 1$ and $K = 3$. Use a majority rule.
Ans. The 1NN classifier gave 39 miles, which is < 50. A 3-NN classifier gives two votes for > 50 and one vote for < 50. Therefore, by majority rule, I should be > 50 miles from home.

3. **Gambling** I am addicted to gambling, and I love playing slot machines. My favorite is a slot machine that, every time I pull the arm, there is a θ chance that I will win \$10, and $(1-\theta)$ chance that I will win nothing. However, each pull costs \$5.

   (a) **(6 pts)** If $\theta = 20\%$, Denote X the number of dollars won at each pull. ($X < 0$ signals I lost some money.) Compute the mean and variance of X.

   Ans. Expectation / mean:

   $$\mathbb{E}[X] = (\$10 - \$5) \cdot \mathbf{Pr}(\text{Win}) + (\$0 - \$5) \cdot \mathbf{Pr}(\text{Lose}) = \$5 \cdot 20\% - \$5 \cdot 80\% = -\$3$$

   Variance:

   $$\mathbb{E}[(X - \mathbb{E}[X])^2] = ((\$10 - \$5) + \$3)^2 \cdot \mathbf{Pr}(\text{Win}) + ((\$0 - \$5) + \$3)^2 \cdot \mathbf{Pr}(\text{Lose}) = 64 \cdot 20\% - 4 \cdot 80\% = 28.8.$$

   (b) **(4 pts)** On Monday, I go in and pull the arm 10 times, and I win 3 times. What is the **maximum likelihood estimate** of θ?

   Ans. As discussed in lecture, the maximum likelihood estimate of θ is k/m, where k is the number of successful pulls, and m is the number of total pulls. This gives $\theta_{\mathbf{MLE}} = 3/10 = 0.3$.

   (c) **(8 pts)** (Tricky) On Tuesday, my friend tells me that she plays this machine all the time, and says "I'm pretty sure that $\theta = 0.7$ for this machine, so play up!" Trouble is, this friend is not that reliable, and tends to overestimate things. I'd say that in fact, the true value of θ follows a uniform distribution with the following probability density function:

   $$p_\theta(\theta) = \begin{cases} 5, & \text{if } 0.5 \le \theta \le 0.7 \\ 0, & \text{else.} \end{cases}$$

   Using this information, and knowing that today I played 10 pulls and won 3 of them, estimate the **maximum a posteriori** estimate of θ.

   Hint: First write down all the formulas you can. When trying to optimize some function, draw some pictures.

   Ans. I will take my own advice and start by writing some equations:

   $$\theta_{\mathbf{MAP}} = \operatorname*{argmax}_\theta \mathbf{Pr}(\theta|D) = \operatorname*{argmax}_\theta \mathbf{Pr}(D|\theta)\mathbf{Pr}(\theta) = \operatorname*{argmax}_\theta \Pr(D|\theta)p_\theta(\theta).$$

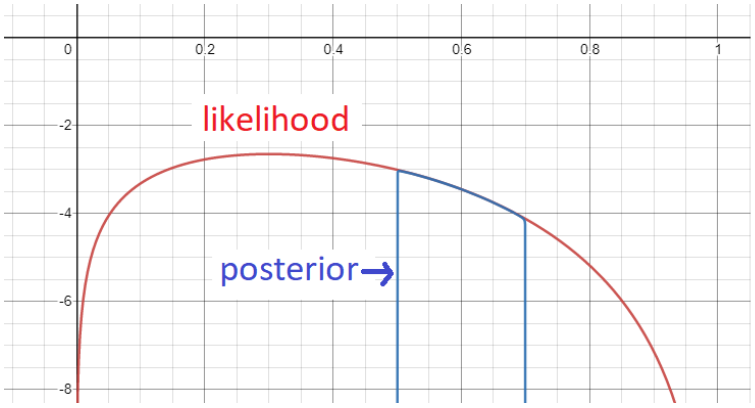
   The likelihood, as used before, is

   $$\mathbf{Pr}(D|\theta) = \binom{k}{m}\theta^k(1-\theta)^{m-k} \propto \theta^3(1-\theta)^7.$$

   Then, for some scaling factor α,

   $$\Pr(D|\theta)p_\theta(\theta) = \begin{cases} \alpha\theta^3(1-\theta)^7 & \text{if } 0.5 \le \theta \le 0.7 \\ 0 & \text{else.} \end{cases}$$

   Well, if I just try plotting $\theta^3(1-\theta)^7$ (or $3\log(\theta) + 7\log(1-\theta)$), I will see that it is a concave curve, with a peak at $\theta = 3/10$, and curving downward from there. (See figure below, red curve.) But, incorporating my prior information, I know that in fact we're not allowed to consider $\theta < 0.5$. (See figure below, blue curve.)



Log of likelihood and posterior (proportional, not exact)

Therefore, the MAP estimate must be

$$\theta_{\mathbf{MAP}} = 0.5.$$

(d) **(12 pts)** On Wednesday, the guy who owns the machines lets slip that $\theta = 0.25$. I decide I will play the machine 4 times. Fill in the table with the likelihoods that I win k out of 4 times. Report numbers to the nearest 0.001 digit.

Ans.

$$\mathbf{Pr}(\text{ win } k \text{ times } |\theta = 0.25) = \binom{m}{k}\theta^k(1-\theta)^{m-k} = \binom{4}{k}0.25^k 0.75^{4-k}$$

k	$\mathbf{Pr}(\text{ I win } k \text{ times } \mid \theta = 0.25)$
0	$1 \cdot 0.3164 \approx 0.316$
1	$4 \cdot 0.1055 \approx 0.422$
2	$6 \cdot 0.0352 \approx 0.211$
3	$4 \cdot 0.0117 \approx 0.047$
4	$1 \cdot 0.0039 \approx 0.004$

What is the maximum likelihood estimate of k, where k is the number of times I win?

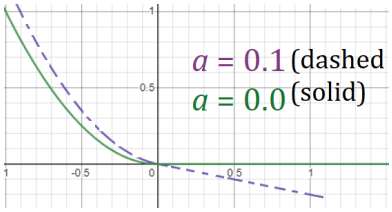
Ans. Based on the table, it is clear that the maximum likelihood estimate is $k = 1$.

4. **Max margin minimization** Consider the following margin maximizing loss function

$$f(\theta) = \frac{1}{m}\sum_{i=1}^{m} g(y_i x_i^T \theta)$$

where g is a smoothed version of a "leaky ReLU" function:

$$g(s) = \begin{cases} -2\alpha s, & s \geq 0 \\ (s-\alpha)^2 - \alpha^2 & s \leq 0. \end{cases}$$



(a) **(3 pts)** Given the minimizer

$$\theta^* = \underset{\theta}{\text{argmin}}\, f(\theta),$$

describe the prediction rule to obtain a label y for a future data feature vector x. Is this classifier a linear classifier, or nonlinear?

Ans. Anything along the lines of $y = \textbf{sign}(x^T\theta^*)$ works. This classifier is linear.

(b) **(3 pts)** Compute the gradient of $f(\theta)$.

Ans. $\nabla f(\theta) = Z^T d$ where

$$d_{[i]} = g'(y_i x_i^T \theta), \quad i = 1, ..., m$$

and

$$Z = \begin{bmatrix} y_1 x_1 & y_2 x_2 & \cdots & y_m x_m \end{bmatrix}$$

and

$$g'(s) = \begin{cases} 2\alpha, & s \geq 0 \\ 2s - 2\alpha & s \leq 0. \end{cases}$$

(c) I get a final solution θ, which has norm $\|\theta\|_2 = 0.5$, and I compute the following quantities in my test dataset (which includes data points x_i and labels y_i): **Ans.**

sample	y_i	$x_i^T\theta$	margin
A	1	0.1	0.2
B	-1	0.1	- 0.2
C	1	-1	-2
D	1	10	20
E	-1	-100	200

i. **(5 pts)** Fill in the 3rd column with the computed margin for each datapoint.

Ans. Remember that margin $= \frac{y_i x_i^T \theta}{\|\theta\|_2}$.

ii. **(2 pts)** Which (if any) points were misclassified?

Ans. Any point where $y_i \neq \textbf{sign}(x_i^T\theta)$ is misclassified. These would be points B and C.

iii. **(2 pts)** Which point was "best classified"? That is, which point is furthest from the decision boundary, on the correctly classified side?

Ans. That would be point E, which has a huge margin of 200.

(d) Show that f is convex. Do this in 2 steps.

i. **(4 pts)** First, show that the following scalar function is convex:

$$q(s) := g(s) + 2\alpha s = (\max\{-s, 0\})^2.$$

Ans. Note that q is not twice differentiable. ($q''(s)$ doesn't exist at $s = 0$.) Therefore, using either the definition of convexity or first order condition works.
In this case, using the first order condition is immensely shorter.

- Pick any u, v.
- Now we look at 4 cases separately. First, suppose that $u \geq 0$ and $v \geq 0$. Then

$$q(u) = q(v) = 0, \qquad q'(u) = q'(v) = 0$$

and thus

$$q(u) - q(v) - q'(v)(u - v) = 0.$$

- Now, suppose that $u \geq 0$ and $v \geq 0$. Then,

$$q(u) - q(v) - q'(v)(u - v) = u^2 - v^2 - 2v(u - v) = u^2 - 2uv + v^2 = (u - v)^2 \geq 0.$$

- Next, suppose that $u \geq 0$, $v \leq 0$. Then since $q(v) = q'(v) = 0$,

$$q(u) - q(v) - q'(v)(v - u) = q(u) = u^2 \geq 0.$$

- Finally, suppose that $u \leq 0$, $v \leq 0$. Then

$$q(u) - q(v) - q'(v)(u - v) = 0 - v^2 - 2v(u - v) = v(v - 2u) \overset{u \leq 0}{\geq} v^2 \geq 0.$$

To use the definition, the math gets considerably harder.

- Pick any u, v. Pick any $0 \geq \beta \geq 1$, and define $w = \beta u + (1 - \beta)v$.
- Again, we look at 4 cases separately. First, suppose that $u \geq 0$ and $v \geq 0$. Then $w \geq 0$, and

$$q(w) = q(u) = q(v) = 0, \qquad q'(u) = q'(v) = 0$$

and thus

$$q(\beta u + (1 - \beta)v) = 0 = \beta q(u) + (1 - \beta)q(v).$$

- Now, suppose that $u \leq 0$ and $v \leq 0$. Then, $w \leq 0$ and

$$q(w) = \beta^2 u^2 + (1 - \beta)^2 v^2 + 2\beta(1 - \beta)uv.$$

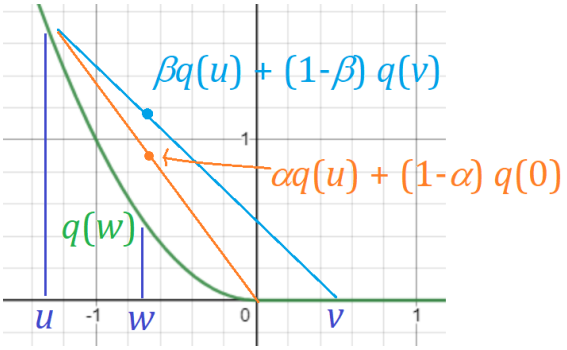
Then

$$\beta q(u) + (1 - \beta)q(v) - q(w) = \beta(1 - \beta)u^2 + (1 - \beta)\beta v^2 - 2\beta(1 - \beta)uv = \underbrace{\beta(1 - \beta)(u - v)^2}_{(*)} \geq 0.$$

- Next, suppose that $u \leq 0 \leq w \leq v$. Then,

$$\beta q(u) + \underbrace{(1 - \beta)q(v)}_{=0} - \underbrace{q(w)}_{=0} = \beta u^2 \geq 0.$$

- Finally, suppose that $u \leq w \leq 0 \leq v$. This one is a bit not straightforward. I will use instead a trick, based on the following picture:



First, observe that for $\alpha = w/u$,

$$w = \alpha u + (1-\alpha)\cdot 0$$

and under our assumptions, $0 \le w/u \le 1$. Our goal will be to show that

$$q(w) \overset{(a)}{\le} \alpha q(u) + (1-\alpha)q(0) \overset{(b)}{\le} \beta q(u) + (1-\beta)q(v).$$

Since $0 \le 0$, (a) follows from the third bullet (condition of $u \le 0$ and $v \le 0$). To show (b),

$$\underbrace{\alpha}_{=w/u} q(u) + (1-\alpha)\underbrace{q(0)}_{=0} = \frac{\beta u + (1-\beta)v}{u} q(u) = \beta q(u) + \frac{(1-\beta)v}{u} q(u) \overset{u\le 0}{\le} \beta q(u) = \beta q(u) + (1-\beta)\underbrace{q(v)}_{=0}$$

- By symmetry of u and v in this case, we can repeat the previous two bullets with u and v switched, thus covering all the cases.

ii. **(4 pts)** Then, use this to argue that f is convex, using properties of convex functions.
Ans. First, consider $f_i(\theta) = g(y_i z_i^T \theta)$
Then, since $g(s)$ is convex in s, then

$$f_i(\beta\theta + (1-\beta)\nu) = g(\beta y_i z_i^T \theta + (1-\beta) y_i z_i^T \nu) \overset{\text{convex } g}{\le} \beta g(y_i z_i^T \theta) + (1-\beta) g(y_i z_i^T \nu) = \beta h(\theta) + (1-\theta) h(\nu).$$

That is, if g is convex then so is f_i.
Working our way to f, notice that for any θ, any ν and any $0 \le \beta \le 1$,

$$f(\beta\theta + (1-\beta)\nu) = \frac{1}{m}\sum_{i=1}^m \underbrace{f_i(\beta\theta + (1-\beta)\nu)}_{\le \beta f_i(\theta) + (1-\beta) f_i(\nu)} \le \beta \frac{1}{m}\sum_{i=1}^m f_i(\theta) + (1-\beta)\frac{1}{m}\sum_{i=1}^m f_i(\nu) = \beta f(\theta) + (1-\beta) f(\nu).$$

Thus, if all f_is are convex, then f must be convex.

(e) **(3 pts)** For $\alpha = 0$ (no leak), argue that if every point x_i is classified correctly, then $f(\theta) = 0$, and $\nabla f(\theta) = 0$.[1]

Ans. Notice that if $\alpha = 0$, then whenever $y_i = \mathbf{sign}(x_i^T\theta)$, then

$$g(y_i x_i^T \theta) = g'(y_i x_i^T \theta) = 0.$$

If this is true for all i, then the function and gradient

$$f(x) = \frac{1}{m}\sum_{i=1}^m \underbrace{g(y_i x_i^T\theta)}_{=0} = 0, \qquad \nabla f(\theta) = \frac{1}{m}\sum_{i=1}^m y_i x_i^T \underbrace{g'(y_i x_i^T\theta)}_{=0} = 0.$$

(f) **(4 pts)** (Hard) For $\alpha > 0$ (leaky), show that if all points are correctly classified, and not all the points are on the margin, then there can be no stationary points; e.g. there does not exist a θ where $\nabla f(\theta) = 0$. In this case, show that gradient descent will produce $\theta^{(t)}$ where $\|\theta^{(t)}\|_2$ never stops increasing as $t \to \infty$.
Ans. In this case, note that if $y_i x_i^T \theta > 0$, then $g'(y_i x_i^T\theta) = 2\alpha > 0$. Now suppose that there exists a point θ where $\nabla f(\theta) = 0$. Then

$$\nabla f(\theta)^T \theta = \frac{1}{m}\sum_{i=1}^m \underbrace{g'(y_i x_i^T\theta)}_{=-2\alpha} \underbrace{y_i x_i^T \theta}_{\ge 0} \overset{\text{at least one } y_i x_i^T\theta > 0}{<} 0.$$

This is not possible if $\nabla f(\theta) = 0$; therefore, there are no points θ where $\nabla f(\theta) = 0$.
For the second part, notice that gradient descent with step size $\eta > 0$ does

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla f(\theta)$$

[1] We call this point an "interpolating point."

Then

$$(\theta^{(t+1)} - \theta^{(t)})^T \theta^{(t)} = \|\theta^{(t)}\|_2^2 - \eta \frac{1}{m} \sum_{i=1}^{m} g'(y_i x_i^T \theta) y_i x_i^T \theta - \|\theta^{(t)}\|_2^2 = \frac{2\eta\alpha}{m} \underbrace{\sum_{i=1}^{m} y_i x_i^T \theta}_{>0} > 0$$

and therefore

$$\|\theta^{(t+1)}\|_2^2 = \|\theta^{(t)}\|_2^2 + \underbrace{2(\theta^{(t+1)} - \theta^{(t)})^T \theta^{(t)}}_{>0} + \underbrace{\|\theta^{(t+1)} - \theta^{(t)}\|_2^2}_{>0}$$

That is to say, the sequence $\|\theta^{(t)}\|_2^2$ just keeps monotonically increasing.