CSCI 567, Spring 2023
Yan Liu

Quiz 2

Released: Mar 7, 2023
Total scores: 100 points

Problem	1	2	3	4	5	Total
Points						

Name: .... **USC ID**: .... **Email**: ....

Instructions

- The exam has a total of **10 pages** (including this cover). Each problem has several questions. Once you are permitted to open your exam (and not before), you should check and make sure that you are not missing any pages.
- Duration of the exam is **90 minutes**. Questions are not ordered by their difficulty. Budget your time on each question carefully. Ask a proctor if you have any question regarding the exam.
- **Select one and only one answer** for all multiple choice questions.
- Answers should be written down **legibly**.
- **You must answer on the space provided for each question**. We make sure that provided spaces are sufficient for the questions we ask. You can use the last blank page as scratch paper.
- **We will not accept any additional page for your answers**.
- This is a **closed-book/notes** exam. Consulting any resources is NOT permitted.
- Any kind of cheating will lead to **score 0** for the entire exam and be reported to SJACS.
- **(For DEN students only)** For all questions, write down your answers as you would normally do for the written assignments, and then submit a pdf/jpg/png for each of the 5 problems. Make sure that your writing is recognizable. Please write down your answers legibly.

Problem 1 Multiple Choice Questions (20 points)

Select **one and only one answer** for all multiple choice questions.

Rubric: correct answer: 2 points, incorrect answer: 0 points

1. Which penalty function should not be used for regularizing model complexity? **(2 points)**

   (a) $R(\mathbf{w}) = \exp\{\sum_i |w_i|\}$

   (b) $R(\mathbf{w}) = \exp\{-\sum_i |w_i|\}$

   (c) $R(\mathbf{w}) = -\sum_i \log(|w_i|^{-1})$

   (d) $R(\mathbf{w}) = \sum_i \exp\{|w_i|\}$

   b: because larger weight $w \rightarrow$ smaller penalty

2. Suppose we are training a neural network with mini-batch SGD of batch size 50, and 50000 training samples. How many updates would there be while training for 5 epochs? **(2 points)**

   (a) 50000

   (b) 1000

   (c) 5000

   (d) 250000

   c: $(50000/50) * 5 = 5000$

3. For $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{2\times 1}$, which of the following bases $\phi(x)$ corresponds to the kernel defined as **(2 points)**

$$k(x, x') = e^{x_1 + x_1'} + e^{2(x_2 + x_2')}$$

   (a) $\phi(x) = [e^{x_1}, e^{x_2}]^T$

   (b) $\phi(x) = [e^{x_1}, \sqrt{2}e^{x_2}]^T$

   (c) $\phi(x) = [e^{x_1}, e^{\sqrt{2}x_2}]^T$

   (d) $\phi(x) = [e^{x_1}, e^{2x_2}]^T$

   d:

4. Consider the dataset consisting of points (x, y), given the basis function $\phi(x, y) = [x^2, 2xy, y^2]^T$, which of the following matrices is the kernel matrix of the three data points $(x_1, y_1) = (1, 0), (x_2, y_2) = (0, 1), (x_3, y_3) = (1, 1)$? **(2 points)**

   (a) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 6 \end{bmatrix}$

   (b) $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 6 \end{bmatrix}$

   (c) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 6 \end{bmatrix}$

   (d) $\begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \\ 2 & 1 & 6 \end{bmatrix}$

c:

5. Which of the following is not a true statement about Lagrangian duality? **(2 points)**

   (a) The solution of primal form and dual form is always equal.
   (b) They can be solved with convex optimization.
   (c) Duality lets us formulate optimality conditions for constrained optimization problems.
   (d) It can be optimized in the dual space.

   a: i.e. weak duality

6. Which of the following on reduction from multiclass classification to binary classification is **incorrect**? **(2 points)**

   (a) One-versus-one is usually more robust than one-versus-all, but is slower in prediction.
   (b) Multiclass logistic regression was invented since its binary version cannot be combined with one-versus-one or other reductions.
   (c) A random code is an option for matrix M in the Error-correcting Output Codes reduction.
   (d) Tree-based reduction is especially useful when the number of possible classes is huge.

   b: logistic regression $(k = 2)$ can certainly be applied to these reductions.

7. Which of the following on SVM is **incorrect**? **(2 points)**

   (a) SVM tries to find a hyperplane with maximum margin, and thus it can only be applied to linearly separable data.
   (b) The primal formulation of SVM minimizes L2 regularized hinge loss.
   (c) Learned weight W is perpendicular to a hyperplane.
   (d) It is possible that a support vector does not satisfy the hard-margin constraint.

   a: Slide lec7-2-2023, Page 33/36

8. For a fixed multiclass problem, which of the following multiclass-to-binary reductions has the smallest testing time complexity? **(2 points)**

   (a) One-versus-all
   (b) One-versus-one
   (c) Tree reduction
   (d) Both (A) and (C)

   c: Slide lec5-2023, Page 23/44

9. We learned that regularization could be used to prevent overfitting. Which technique could also be used to prevent overfitting in neural nets? **(2 points)**

   (a) Retraining on the same data many times.
   (b) Using a validation set for early stopping.
   (c) Dropping random neurons in each iteration of backpropagation.
   (d) Both (B) and (C)

   d: Slide lec5-2023, Page 39/44

10. Which of the following is **incorrect**? **(2 points)**

(a) A machine learning algorithm can be kernelized if it only uses the feature vectors through their inner products.

(b) A Gram/kernel matrix must be positive semidefinite.

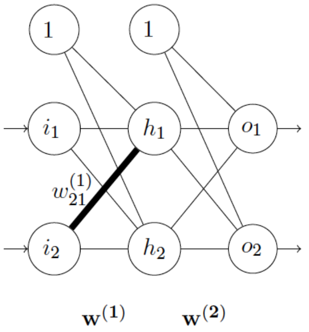
(c) SVM cannot be solved by stochastic gradient descent.

(d) Neural nets with a linear activation function is a linear function.

c: Slide lec7-2-2023, Page 16/36

Problem 2 Neural Networks (20 points)

A neural network with two inputs, two hidden neurons with biases, and two output neurons is given:



Let $w_{21}^{(1)}$ be the weight assigned to the connection between input neuron i_2 and hidden neuron h_1. The loss function is $L = \frac{1}{2}\sum_{i=1}^{2}(y_i - o_i)^2$ where

z

$$o_i = \sigma(w_{1i}^{(2)}h_1 + w_{2i}^{(2)}h_2 + b_i^{(2)})$$

$$h_i = \sigma(w_{1i}^{(1)}i_1 + w_{2i}^{(1)}i_2 + b_i^{(1)})$$

and $\sigma(.)$ is the sigmoid function. Note that $\sigma(z) = \frac{1}{1+\exp(-z)}$. Follow the steps below to obtain $\frac{\partial L}{\partial w_{21}^{(1)}}$.

(a) Prove that $\frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1-\sigma(z))$. **(4 points)**

$$\begin{aligned}\frac{\partial \sigma(z)}{\partial z} &= \frac{\partial}{\partial z}\left(\frac{1}{1+\exp(-z)}\right)\\ &= \frac{-1}{(1+\exp(-z))^2}\exp(-z)(-1)\\ &= \frac{1}{1+\exp(-z)}\cdot\frac{\exp(-z)}{1+\exp(-z)}\\ &= \sigma(z)(1-\sigma(z))\end{aligned}$$

Rubric: Give partial credit of 2 points for partially correct proofs, otherwise full credit.

(b) Calculate $\frac{\partial L}{\partial o_1}$, $\frac{\partial o_1}{\partial h_1}$ and $\frac{\partial h_1}{\partial w_{21}^{(1)}}$. **(8 points)**

$$\frac{\partial L}{\partial o_1} = o_1 - y_1 \quad \textbf{(2 points)}$$
$$\frac{\partial o_1}{\partial h_1} = o_1(1-o_1)w_{11}^{(2)} \quad \textbf{(3 points)}$$
$$\frac{\partial h_1}{\partial w_{21}^{(1)}} = h_1(1-h_1)i_2 \quad \textbf{(3 points)}$$

do/dh = (do/dz) * (dz/dh)
do/dz = sig(z)(1 - sig(z))
dz/dh_1 = w_11
do/dh = o_1(1 - 0_1)w_11

Rubric: Give half credit (per sub score) for partial mistakes. For example, setting up the correct chain rule of $\frac{\partial o_1}{\partial h_1}$ but get the wrong derivative result—students will get 1.5 points for the second sub score.

(c) Using either the symmetry of the network or by explicitly computing them, calculate $\frac{\partial L}{\partial o_2}$ and $\frac{\partial o_2}{\partial h_1}$ **(4 points)**

$$\frac{\partial L}{\partial o_2} = o_2 - y_2 \quad \textbf{(2 points)}$$
$$\frac{\partial o_2}{\partial h_1} = o_2(1-o_2)w_{12}^{(2)} \quad \textbf{(2 points)}$$

Rubric: Give half credit (per sub score) for partial mistakes either the mistake(s) were made in the previous sub question or in this question.

(d) Use the previous results and the chain rule of calculus to find $\frac{\partial L}{\partial w_{21}^{(1)}}$. **(4 points)**

$$\begin{aligned}\frac{\partial L}{\partial w_{21}^{(1)}} &= \left(\frac{\partial L}{\partial o_1}\frac{\partial o_1}{\partial h_1} + \frac{\partial L}{\partial o_2}\frac{\partial o_2}{\partial h_1}\right)\frac{\partial h_1}{\partial w_{21}^{(1)}} \quad \textbf{(2 points)}\\ &= \left((o_1-y_1)o_1(1-o_1)w_{11}^{(2)} + (o_2-y_2)o_2(1-o_2)w_{12}^{(2)}\right)h_1(1-h_1)i_2 \quad \textbf{(2 points)}\end{aligned}$$

Rubric: Give half credit (per sub score) for partial mistakes (like correct chain rule but wrong derivative substituted).

Problem 3 Logistic Regression (20 points)

Given a dataset $D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_n)\}$ where $x_n \in \mathbb{R}^D$, and $y_n \in \{0, 1\}$. Consider this prediction model

$$P(y_n = 1|x_n; w) = \Phi(w^T x_n),$$

where

$$\Phi(z) = \int_{-\infty}^{z} \phi(v)dv,$$

and

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

The shape of $\Phi(z)$ is very similar to the sigmoid activation function we used in logistic regression. Because $\Phi(z)$ is called the *probit function*, we thus call this model *Probit Regression*.

The cross-entropy loss of a binary classifier over a dataset is defined as follows:

$$H(y,p) := -\sum_{(x_n,y_n)\in D} \big(y_n \ln p_n + (1-y_n)\ln(1-p_n)\big),$$

where $p_n = P(y_n = 1|x_n; w)$. Our goal is to minimize cross-entropy in our binary classification problem. Please derive $\nabla_w H(y, P(y|x,w))$, express it with $y_n, x_n, \Phi(\cdot)$ and $\phi(\cdot)$, and reduce it to the simplest form.

Hint Given $\frac{\partial \Phi(z)}{\partial z} = \phi(z)$, applying the chain rule to get the derivation w.r.t. w.

(a) Calculate $\frac{\partial}{\partial w}(y_n \ln p_n)$ **(5 points)**

Denote $w^T x_n$ by z_n. Denoting logarithm by ln or log are both fine.

$$\frac{\partial}{\partial w} z_n = \frac{\partial}{\partial w} w^T x_n = x_n$$

$$\frac{\partial}{\partial z_n}\Phi(z_n) = \phi(z_n)$$

$$\frac{\partial}{\partial \Phi}\ln\Phi(z_n) = \frac{1}{\Phi(z_n)}$$

$$\frac{\partial}{\partial z}\ln\Phi(z_n) = \left(\frac{\partial}{\partial \Phi(z_n)} ln\Phi(z_n)\right)\times\left(\frac{\partial}{\partial z_n}\Phi(z_n)\right) = \frac{\phi(z_n)}{\Phi(z_n)}$$

$$\frac{\partial}{\partial w}\ln\Phi(z_n) = \left(\frac{\partial}{\partial z_n}\ln\Phi(z_n)\right)\times\left(\frac{\partial}{\partial w} z_n\right) = \frac{\phi(z_n)}{\Phi(z_n)} x_n$$

(3 points)

$$g(x) = y_n, \;\; h(x) = \ln p_n$$

$$\frac{\partial}{\partial w}(y_n \ln p_n) = \frac{\partial}{\partial w} g(x)h(x) = g'(x)h(x) + g(x)h'(x)$$

$$g'(x) = 0, \;\; h'(x) = \frac{\phi(z_n)}{\Phi(z_n)} x_n$$

$$\frac{\partial}{\partial w}(y_n \ln p_n) = \frac{\phi(z_n)}{\Phi(z_n)} y_n x_n$$

(2 points)

Rubric: Give half credit (per sub score) for partial mistakes (like correct chain rule but wrong derivative substituted either the wrong substitution comes from using previous question's answer(s) or current question's value(s)).

(b) Calculate $\frac{\partial}{\partial w}((1-y_n)\ln(1-p_n))$ **(5 points)**

$$\frac{\partial}{\partial \Phi}\ln(1-\Phi(z_n)) = \frac{-1}{1-\Phi(z_n)}$$

$$\frac{\partial}{\partial z_n}\ln(1-\Phi(z_n)) = \frac{-\phi(z_n)}{1-\Phi(z_n)}$$

$$\frac{\partial}{\partial w}\ln(1-\Phi(z_n)) = \frac{-\phi(z_n)}{1-\Phi(z_n)} x_n$$

(3 points)

$$g(x) = 1 - y_n, \;\; h(x) = \ln(1-p_n)$$

$$\frac{\partial}{\partial w}((1-y_n)\ln(1-p_n)) = \frac{\partial}{\partial w} g(x)h(x) = g'(x)h(x) + g(x)h'(x)$$

$$g'(x) = 0, \;\; h'(x) = \frac{-\phi(z_n)}{1-\Phi(z_n)} x_n$$

$$\frac{\partial}{\partial w}((1-y_n)\ln(1-p_n)) = (1-y_n)\frac{-\phi(z_n)}{1-\Phi(z_n)} x_n$$

(2 points)

Rubric: Give half credit (per sub score) for partial mistakes (like correct chain rule but wrong derivative substituted either the wrong substitution comes from using previous question's answer(s) or current question's value(s)).

(c) Calculate $\nabla_w H(y, P(y|x,w))$ **(2 points)**

$$\Rightarrow \nabla_w H(y, P(y|x,w)) = -\sum_{n=1}^{N} \left(\frac{y_n}{\Phi(w^T x_n)} - \frac{1-y_n}{1-\Phi(w^T x_n)}\right)\phi(w^T x_n) x_n$$

$$= \sum_{n=1}^{N} \frac{\Phi(w^T x_n) - y_n}{\Phi(w^T x_n)(1-\Phi(w^T x_n))}\phi(w^T x_n) x_n$$

(2 points)

Rubric: If students apply the chain rule correctly throughout the (a), (b) and (c) but unfortunately there was one (or a few) computational mistakes that cause the final result to not be exactly correct, give students 8 points (from 12 points).

Rubric: if students do not substitute the final results to be represented with $w, y_n, x_n, \Phi(\cdot)$ and $\phi(\cdot)$, deduct 1 point.

Outliers are non-typical data points that deviates far away from typical ones with the same label. For example, if the data points $x_n \in \mathcal{R}$ with label $y_n = 1$ are mostly within the range $[2, 4]$, then a data point with a value 10 is considered to be an outlier. Suppose that we add an outlier x_{N+1} to the dataset D.

(d) Calculate $\nabla_{x_{N+1}} H(y, P(y|x,w))$ **(8 points)**

Note that the cross-entropy is now a summation over $N+1$ points, i.e.,

$$H(y,p) := -\sum_{n=1}^{N+1} \left(y_n \ln p_n + (1-y_n)\ln(1-p_n)\right),$$

Hint Following the same steps you did on (a), (b), (c) to solve (d). Remember this time you will take derivative w.r.t. x_{N+1}.

$\nabla_{x_{N+1}} w^T x_{N+1} = w$ **(1 points)**

$\nabla_{z_n} \ln \Phi(z_n) = \frac{1}{\Phi(z_n)}\phi(z_n) = \frac{\phi(z_n)}{\Phi(z_n)}$ **(1 points)**

$\nabla_{z_n} \ln(1-\Phi(z_n)) = \frac{1}{1-\Phi(z_n)}(-\phi(z_n)) = \frac{-\phi(z_n)}{1-\Phi(z_n)}$ **(1 points)**

$$
\begin{aligned}
\nabla_{x_{N+1}} H(y,p) &= \nabla_{x_{N+1}} - \sum_{n=1}^{N+1} \left(y_n \ln p_n + (1-y_n)\ln(1-p_n)\right) \\
&= -\nabla_{x_{N+1}} \left(y_n \ln \Phi(z_{N+1}) + (1-y_n)\ln(1-\Phi(z_{N+1}))\right) && \textbf{(2 points)} \\
&= -\left(y_n \nabla_{x_{N+1}} \ln \Phi(z_{N+1}) + (1-y_n)\nabla_{x_{N+1}} \ln(1-\Phi(z_{N+1}))\right) \\
&= -(y_n \frac{\phi(z_n)}{\Phi(z_n)} w + (1-y_n)(\frac{-\phi(z_n)}{1-\Phi(z_n)})w) && \textbf{(2 points)} \\
&= -(\frac{y_n}{\Phi(z_n)}\phi(z_n)w - \frac{1-y_n}{1-\Phi(z_n)}\phi(z_n)w) \\
&= \frac{\Phi(z_n) - y_{N+1}}{\Phi(z_n)(1-\Phi(z_n))}\phi(z_n)w && \textbf{(1 points)}
\end{aligned}
$$

where $z_n = w^T x_{N+1}$. The derivations are similar to Problem (a), (b) and (c). Note that the summation is gone because we are taking derivative w.r.t a single data point.

Rubric: If students apply the chain rule correctly throughout the question but unfortunately there was one (or a few) computational mistakes that cause the final result to not be exactly correct, give students 6 points (from 8 points).

Rubric: For question (d), students do not need to convert z_n to $w^T x_{N+1}$ in the final result. If students did do the substitution, great! if not, do not deduct the score.

Problem 4 SVM and Lagrangian Duality (20 points)

Consider a dataset consisting of points in the form of (x, y), where x is a real value, and $y \in \{-1, 1\}$ is the label. There are only three points $(x_1, y_1) = (-1, -1), (x_2, y_2) = (2, -1), (x_3, y_3) = (0, 1)$.

(a) Given the feature mapping $\phi(x) = [x, x^2]^T$, write down the corresponding kernel function $k(x, x')$. **(2 points)**

$$k(x, x') = xx' + (xx')^2$$

(b) Write down the Gram matrix of this dataset. **(3 points)**

$$\begin{bmatrix} 2 & 2 & 0 \\ 2 & 20 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Rubric: Deduct 1 point for ≤ 2 computational mistakes, Deduct 2 point for 3-4 computational mistakes, > 4 mistakes 0 points.

(c) Given the dual formulation of SVM below (with the hyperparmeter $C = +\infty$).

$$\max_{\alpha} \quad \sum_n \alpha_n - \frac{1}{2} \sum_{m,n} y_m y_n \alpha_m \alpha_n k(x_m, x_n)$$

$$s.t. \quad \alpha_n \geq 0, \forall n \quad and \quad \sum_n \alpha_n y_n = 0.$$

Use the Gram matrix in (b) to plug in this dataset into the dual formation. **(5 points)**

Plugging in the specific dataset gives:

$$\max_{\alpha_1, \alpha_2, \alpha_3 \geq 0} \quad \alpha_1 + \alpha_2 + \alpha_3 - \alpha_1^2 - 10\alpha_2^2 - 2\alpha_1\alpha_2 \quad \textbf{(4 points)}$$

$$s.t. \quad \alpha_1 + \alpha_2 = \alpha_3 \quad \textbf{(1 points)}$$

Rubric: Deduct x point for x computational mistakes, where $x \in [1, 5]$, the mistake could be from this subproblem itself or previous subproblem(s).

(d) Solve the dual formulation and state which points (α_n^*) are support vectors. **(5 points)**

Note show your derivation

Substituting α_3 with $\alpha_1 + \alpha_2$, the objective becomes

$$2\alpha_1 + 2\alpha_2 - \alpha_1^2 - 10\alpha_2^2 - 2\alpha_1\alpha_2 \quad \textbf{(1 points)}$$

Setting the gradient to zero gives

$$2 - 2\alpha_1 - 2\alpha_2 = 0 \quad \textbf{(1 points)}$$

$$2 - 20\alpha_2 - 2\alpha_1 = 0 \quad \textbf{(1 points)}$$

Solving these linear equations gives $\alpha_1^* = 1, \alpha_2^* = 0$ and $\alpha_3^* = \alpha_1^* + \alpha_2^* = 1$. **(1 points)**

Therefore, only x_1 and x_3 are support vectors. **(1 points)**

(e) Write down the primal solution $w^* \in \mathbb{R}^2$ and $b^* \in \mathbb{R}$, given that

$$w^* = \sum_{n=1}^{3} y_n \alpha_n^* \phi(x_n)$$
$$b^* = y_n - \sum_m y_m \alpha_m^* k(x_m, x_n)$$

(4 points)

$$w^* = (1, -1) \quad \textbf{(2 points)}$$
$$b^* = 1 \quad \textbf{(2 points)}$$

Rubric: Deduct x point for x computational mistakes, where $x \in [1, 4]$, the mistake could be from this subproblem itself or previous subproblem(s).

(f) Given the primal solution of this SVM, what is the prediction of a test point $x = 2$? **(1 points)**

$$SGN({w^*}^T \phi(x) + b^*) = SGN([1, -1]^T [2, 4] + 1) = SGN(-1) = -1$$

Problem 5 Kernel Methods (20 points)

We will explore compositions of kernels to establish new valid kernels. Please use the below properties to prove the validity of these new kernels.

1. $k(x, y) = a_1 k_1(x, y) + a_2 k_2(x, y)$ is a valid kernel, if $a_1, a_2 \geq 0$ and $k_1(\cdot, \cdot)$ and $k_2(\cdot, \cdot)$ are valid kernels.
2. $k(x, y) = k_1(x, y) k_2(x, y)$ is a valid kernel, if $k_1(\cdot, \cdot)$ and $k_2(\cdot, \cdot)$ are valid kernels.

(a) Suppose $k(\cdot, \cdot)$ is a valid kernel, prove that $k_3(x, y) = g(k(x, y))$ is also a valid kernel, where $g(\cdot)$ is a polynomial function with positive coefficients. **(10 points)**

Note that results 1. and 2. also apply to weighted sum and product of more than two kernels by induction. Let $k_3(x, y) = g(k(x, y)) = \sum_{i=0}^{d} a_i k^i(x, y)$ be a polynomial of degree d with $a_i \geq 0 \ \forall i$. Hence, each term of the form $k^i(x, y)$ is a valid kernel using 1 **(5 points)** and the linear combination of all powers with non-negative coefficients is a valid kernel using 2 **(5 points)**.

(b) Suppose $k(\cdot, \cdot)$ is a valid kernel, prove that $k_4(x, y) = \exp(k(x, y))$ is also a valid kernel. **(10 points)**

Hint Think of Taylor series: $exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$

Convert k_4 to Taylor series **(5 points)**

This is a polynomial of infinite degree with all positive coefficients, hence using result of previous part k_4 is a valid kernel. **(5 points)**