Machine Learning
Fall 2023 Final Exam
Date: 15/1/2024
Duration: 100 minutes

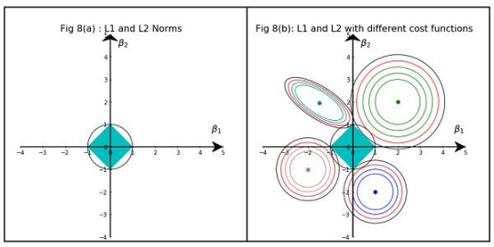
Student Number: _____

Name & Signature: _____

1. (15 points) In regularized regression, we find an optimal set of parameters as follows:

$$\hat{w}_r = \arg\min_w \sum_{i=1}^{n} (y_i - x_i^T w)^2 + \lambda r(W) \tag{1}$$

where $r(w) = ||w||_2^2 = \sum_{j=1}^{d} |w_j|^2$ for squared L_2-norm (a.k.a Ridge) penalty and $r(w) = ||w||_1 = \sum_{j=1}^{d} |w_j|$ for L_1-norm (a.k.a Lasso) penalty. Some possible loss (a.k.a cost) surfaces are shown on the right for the two penalty terms on the right figure below.



Let's also define L_{inf}-norm penalty as $r(w) = \max_i |w_i|$ and $L_{\frac{1}{2}}$-norm penalty as $r(w) = ||w||_{\frac{1}{2}} = \sum_{j=1}^{d} \sqrt{|w_j|}$. Draw two more regions induced by these two new norms on the left figure.

2. (10 points) In HW 2, we considered the MNIST dataset, but for binary classification. Specifically, the task was to determine whether a digit was a 2 or 7. There, we let $Y = 1$ for all the "7" digits in the dataset, and used $Y = 1$ for "2". We used regularized logistic regression. Given a binary classification dataset $\{(x_i, y_i)\}_{i=1}^{n}$ for $x_i \in R^d$ and $y_i \in \{1, 1\}$ we showed in class that the regularized negative log likelihood objective function can be written as

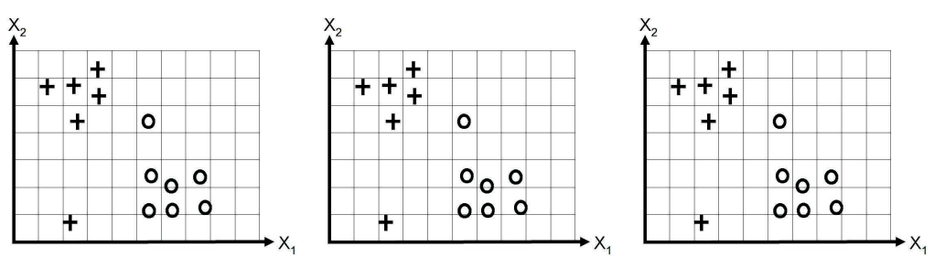
$$J(w, b) = \frac{1}{n} \sum_{1}^{n} log(1 + exp(-y_i(b + x_i^T w))) + \lambda ||w||_2^2 \tag{2}$$

   (a) After you learned the parameters, according to which rule you classified the new points?

   (b) For both the training set and the test, plot how $J(w, b)$ looked as a function of the iteration number (and show both curves on the same plot).

3. (15 points) Consider the data $(x_i = (x_1^{(i)}, x_2^{(i)}), y^{(i)})_{i=1}^{n}$ below, where we fit the model $p(y = 1|x, w) = (w_0 + w_1 x_1 + w_2 x_2)$. Suppose we fit the model by the Lasso regularized negative log-likelihood objective function, i.e., given $\tilde{x}^{(i)} = (1, x^{(i)})$, we minimize

$$J(w, b) = \frac{1}{n} \sum_{1}^{n} log(1 + exp(-y^{(i)}(\tilde{x}^{(i)T} w))) + \lambda \sum_{j=1}^{d} |w_j| \tag{3}$$



   (a) Draw a possible decision boundary corresponding to the optimal model on the leftmost figure.

   (b) Now suppose we regularize only the w_0 parameter, i.e., we minimize

$$J(w, b) = \frac{1}{n} \sum_{1}^{n} log(1 + exp(-y^{(i)}(\tilde{x}^{(i)T} w))) + \lambda |w_0| \tag{4}$$

   Suppose λ is a very large number, so we regularize w_0 all the way to 0, but all other parameters are unregularized. Draw a possible decision boundary on the middle figure.

   (c) Now similarly suppose we regularize only the w_2 parameter. Draw a possible decision boundary on the righmost figure.

4. (2.5 points) True or False: The training error of a function on the training set provides a good estimate of the true error of that function.

5. (2.5 points) True or False: If the step-size for gradient descent is too large, it may not converge.

6. (2.5 points) True or False: We use non-linear activation functions in a neural network's hidden layers so that the network learns non-linear decision boundaries.

7. (2.5 points) True or False: Given a neural network, the time complexity of the backward pass step in the backpropagation algorithm can be prohibitively larger compared to the relatively low time complexity of the forward pass step.

8. (10 points) Suppose you are given the following training inputs $X = [-4; 5; 3]$ and $Y = [-10; 12; 8]$. We want to find an optimal parameter w using Lasso regression as follows

$$\hat{w}_r = \arg\min_w \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T w)^2 + \lambda |w| \tag{5}$$

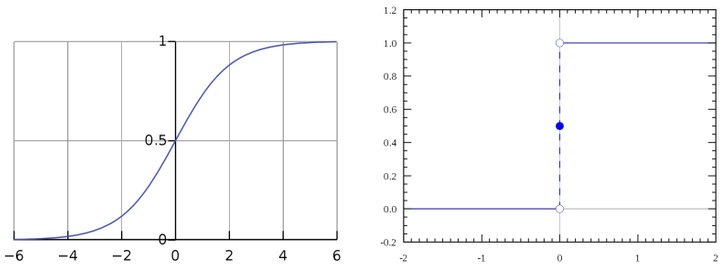
Note that there is no other parameter than w in the model. Assuming $\lambda = 10$, which of the following is closest to the optimal ridge regression parameter? Show your work briefly.

(a) 3.0 **(b)** 4.0 **(c)** 5.0 **(d)** 6.0

9. (15 points) In Homework 2, you have implemented a logistic regression model for binary classification

$$P(Y = y|x, w) = \sigma(y(w^T x + b)) = \frac{1}{1 + \exp(-y(w^T x + b))} \tag{6}$$

where the sigmoid function σ has the shape on the left figure



You decide to train your model using gradient descent. However, before you can turn your assignment in, your friends stop by to give you some suggestions.

(a) Lennie sees your regressor implementation and says there's a much simpler way! Just leave out sigmoids, and let $y = w^T x + b$. The derivatives are a lot less hassle and it runs faster. What's wrong with Lennie's approach?

(b) George comes in and says that your logistic regressor is too complicated, but for a different reason. The sigmoids are confusing and basically the same answer would be computed if we used sign functions, on the right figure above, instead of the sigmoid. What's wrong with George's approach?

10. (10 points) Suppose we have a Ridge regularized linear regression model: $\hat{w} = \arg\min_w \sum_{i=1}^{n} (y_i - x_i^T w)^2 + \lambda ||w||_2^2$. What is the effect of decreasing λ on bias and variance?

11. (15 points) Suppose we want to compute 4-Fold Cross-Validation error on 200 training examples. We need to train a model and compute its error N_1 times, and the Cross-Validation error is the average of the errors. To compute each error, we need to build a model with data of size N_2, and test the model on the data of size N_3. Guess appropriate numbers for N_1, N_2 and N_3?