Data Science (CSM 148)
March 17, 2025

Final Exam

Duration: 120 minutes

Write your name and UID:

Note 1: If you find a question difficult, move on with the rest of the questions and come back to it in the end!

Note 2: Your final grade will be curved, if necessary.

Note 3: There are 7 questions. Only answers written in the boxes will be graded. If you need extra space, please ask for an extra sheet. Good Luck! :-)

1 Linear Regression & Bias: True or False (22 Points)

White 'T' or 'F' in the box corresponding to each of the statements below.

1. **(2 points)** [F] In a linear regression model, a large value for β and $p-$value < 0.05 shows that the corresponding predictor causes the response.
2. **(2 points)** [T] A confounding factor causes a high correlation between multiple other predictors and the response.
3. **(2 points)** [T] If R^2 is large and the residual plot has a normal distribution centered at zero, the model complexity is enough for the given data.
4. **(2 points)** [F] A relatively small value for R^2 means that increasing the model complexity improves the performance.
5. **(2 points)** [F] If the 95% confidence interval for a predictor is [-0.2, 100], there is a significant relationship between the predictor and the response.
6. **(2 points)** [T] Multicollinearity doesn't affect the performance of the model, but makes interpretation of coefficients unreliable.
7. **(2 points)** [T] Stratified sampling can reduce bias of data collection, when the population is imbalanced.
8. **(2 points)** [T] Any type of regularization with appropriate regularization coefficient yields a higher training error but a lower validation error.
9. **(2 points)** [T] Lasso (L1 regularization) makes some coefficients exactly equal to zero and can make interpretation easier.
10. **(2 points)** [T] Larger dataset yields a smaller confidence interval for the predictors, compared to smaller dataset coming from the same distribution.
11. **(2 points)** [T] Mini-batch stochastic gradient descent with appropriate batch size can train a neural network to a better performance compared to gradient descent.

2 Logistic Regression (18 Points)

Fig. 1 shows the log odds of success, based on X_1 (real-valued) and X_2 (categorical with 3 categories {C1, C2, C3}). We use binary variable(s) to model X_2 and model C3 by making all binary variable(s) equal to 0.

Write the logistic model corresponding to Fig. 1. This part has 0 credit, but we won't grade the subsequent parts of this question if your logistic model is missing.

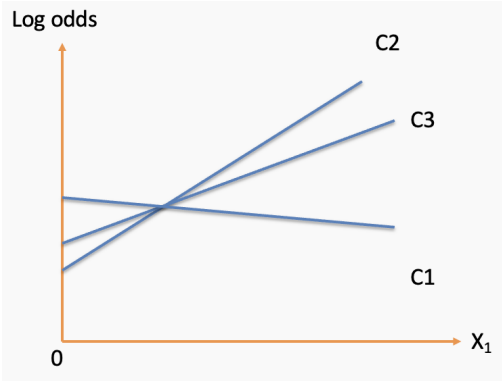


Figure 1: Log odds of success

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_{21} + \beta_3 X_{22} + \beta_4 X_1 X_{21} + \beta_5 X_1 X_{22}$$

Answer the following questions based on Fig. 1 and your logistic model. For parts (d), (e), (f), choose one of the options (from 1 to 10) at the end of the question.

(a) **(2 points)** How many binary variables do you use for X_2 to have an interpretable logistic regression model? 2

(b) **(2 points)** How many interaction terms are used in the logistic regression model corresponding to Fig. 1? 2

(c) **(2 points)** Write the sign of the coefficient of the interaction term(s). If you have more than 1 interaction term, write the sign for all (ordering doesn't matter). -,+

(d) **(3 points)** What's the interpretation of β_i, where β_i is the intercept of the logistic model? 3

(e) **(3 points)** What's the interpretation of β_i, where β_i is the coefficient of X_1? 1

(f) **(3 points)** What's the interpretation of the β_i, where β_i is coefficient of a binary variable used for modeling X_2? 7

(g) **(3 points)** What's the interpretation of the β_i, where β_i is coefficient of an interaction term between X_1 and a binary variable used for modeling X_2? 6

1. One unit increase in X_1 changes the odds of success for C3 by $(e^{\beta_i} - 1)\%$.
2. One unit increase in X_1 changes the odds of success for C3 by β_i.
3. e^{β_i} is the odds of success at $X_1 = 0$ for C3.
4. e^{β_i} is the odds of success for C3.

5. β_i is the odds of success for C3.
6. For one unit increase in X_1, e^{β_i} is the ratio of the multiplicative change in odds of success for C1 or C2 over that of C3.
7. At $X_1 = 0$, e^{β_i} is the ratio of the multiplicative change in odds of success for C1 or C2 over that of C3.
8. At $X_1 = 0$, β_i is the additional change in odds of success for C1 or C2 over that of C3.
9. Odds ratio of success for C1 or C2 over that of C3 at $X_1 = 0$.
10. Odds ratio of success for C1 or C2 over that of C3.

3 Classification Metrics (10 Points)

For a trained binary classifier, we sort examples based on their probability of having label 1.

White 'T' or 'F' in the box corresponding to each of the statements below.

Hint: AUC stands for Area Under the RoC Curve.

c**(2 points)** [F] If the classifier sorts the examples correctly, the model has a good accuracy. **(2 points)** [T] If the classifier sorts the examples correctly, the model has a good AUC. **(2 points)** [F] If the classifier sorts the examples correctly, the model has a good accuracy and a good AUC. **(2 points)** [T] If the model has a good AUC but a poor accuracy for class 1, changing the threshold for predicting class 1 improves the accuracy. **(2 points)** [F] If the model has a good accuracy for class 1 but a poor AUC, changing the threshold for predicting class 1 improves the accuracy.

4 Multi-class Logistic Regression (8 Points)

(a) **(4 points)** Fig. 2 shows the binary classifiers used for multinomial and One-vs-Rest (OvR) Logistic Regressions. Remember that for one of the logistic models, one of the classifiers is derived from the other two. Which figure corresponds to multinomial and OvR models? Write (a) or (b) in the boxes below:

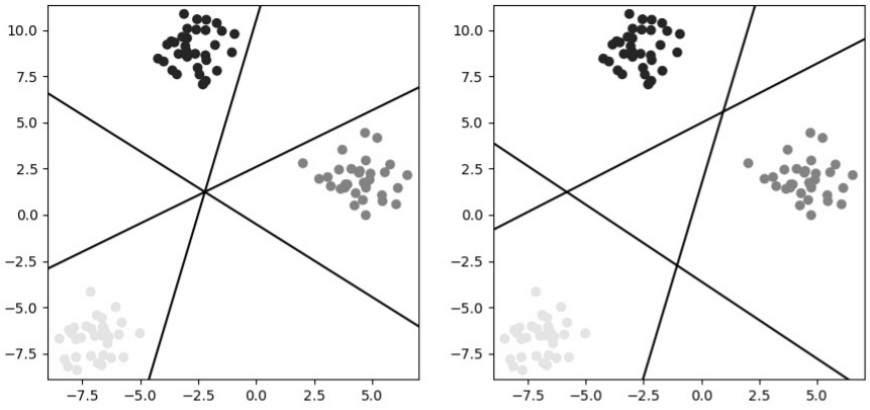
1. Multinomial [a]
2. OVR [b]

(b) **(4 points)** We want to do greedy (forward) model selection (greedy step-wise variable selection) for *each of the linear classifiers* in a multinomial or OvR logistic regression. If we have c classes, and p predictors, how many logistic regression models should we fit in total to choose $q \leq p$ predictors for each model? Assume, we only consider the original predictors in the model selection without including any higher order terms. In each of the boxes below, choose one option from (a) to (f) at the end of the question.

Multinomial: [F]

OvR: [E]

(A) $c \times [p! + 1]$
(B) $(c-1) \times [p! + 1]$
(C) $c \times [\frac{p!}{q! \times (p-q)!} + 1]$
(D) $(c-1) \times [\frac{p!}{q! \times (p-q)!} + 1]$



(a) (b)

Figure 2: Multiclass Logistic Regression

(E) $c \times [(p + p - 1 + \cdots + p - q + 1) + 1]$

(F) $(c - 1) \times [(p + p - 1 + \cdots + p - q + 1) + 1]$

5 Neural Networks (17 Points)

Consider the following neural network with sigmoid activation functions in hidden nodes h_1, h_2, and sigmoid output p.
Remember: $\frac{\partial \log z}{\partial z} = \frac{1}{z}$, and for a sigmoid function p, we have $\frac{\partial p(z)}{\partial z} = p(1 - p)$.

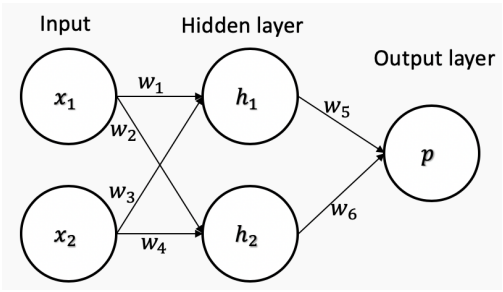


Figure 3: Neural Network

(a) **(3 points)** Consider a mean squared loss, i.e. $\mathcal{L} = \frac{1}{2}\sum_i (p_i - y_i)^2$. Use the chain rule to write the derivative of the loss w.r.t. w_6.

(A) $\frac{\partial \mathcal{L}}{\partial w_6} = (p - y)h_2$

(B) $\frac{\partial \mathcal{L}}{\partial w_6} = (p - y)p(1 - p)h_2$

(C) $\frac{\partial \mathcal{L}}{\partial w_6} = 2(p - y)ph_2$

(D) $\frac{\partial \mathcal{L}}{\partial w_6} = 2(p - y)(1 - p)h_2$

(E) $\frac{\partial \mathcal{L}}{\partial w_6} = (p - y)p(1 - p)$

Answer: B

(b) **(5 points)** Consider a binary cross entropy loss: $\mathcal{L} = \sum_i y_i \log p_i + (1 - y_i) \log(1 - p_i)$. Use the chain rule to write the derivative of the loss w.r.t. w_2.

(A) $\frac{\partial \mathcal{L}}{\partial w_2} = (\frac{y}{p} + \frac{1-y}{1-p})p(1-p)w_6 h_2(1-h_2)x_1$

(B) $\frac{\partial L}{\partial w_2} = (\frac{y}{p} - \frac{1-y}{1-p})p(1-p)h_2(1-h_2)x_1$

(C) $\frac{\partial \mathcal{L}}{\partial w_2} = (\frac{y}{p} - \frac{1-y}{1-p})p(1-p)w_6 x_1$

(D) $\frac{\partial \mathcal{L}}{\partial w_2} = (\frac{y}{p} - \frac{1-y}{1-p})p(1-p)w_6 h_2(1-h_2)x_1$

(E) $\frac{\partial \mathcal{L}}{\partial w_2} = (\frac{y}{p} - \frac{1-y}{1-p})p(1-p)w_6 h_2 x_1$

(F) $\frac{\partial \mathcal{L}}{\partial w_2} = (\frac{y}{p} - \frac{1-y}{1-p})p(1-p)h_2(1-h_2)x_1$

(G) $\frac{\partial \mathcal{L}}{\partial w_2} = (\frac{y}{p} - \frac{1-y}{1-p})p(1-p)h_2(1-h_2)$

Answer: D

(c) **(3 points)** Write one step of *mini-batch stochastic gradient descent* update for w_2, using η as learning rate. Assume there are k examples in each mini-batch.

(A) For $i \in \{1, \cdots k\}$ do $w_2 = w_2 - \eta \frac{\partial \mathcal{L}_i}{\partial w_2}$

(B) $w_2 = w_2 - \frac{\eta}{k}(\frac{\partial \mathcal{L}_1}{\partial w_2} + \frac{\partial \mathcal{L}_2}{\partial w_2} + \cdots + \frac{\partial L_k}{\partial w_2})$

(C) For $i \in \{1, \cdots k\}$ do $w_2 = w_2 - \frac{\eta}{k}(\frac{\partial \mathcal{L}_1}{\partial w_2} + \frac{\partial \mathcal{L}_2}{\partial w_2} + \cdots + \frac{\partial L_k}{\partial w_2})$

(D) For $i \in \{1, \cdots k\}$ do $w_2 = w_2 - \frac{\eta}{n}(\frac{\partial \mathcal{L}_1}{\partial w_2} + \frac{\partial \mathcal{L}_2}{\partial w_2} + \cdots + \frac{\partial \mathcal{L}_n}{\partial w_2})$

Answer: B

(d) **(2 points)** Assuming the data has a pattern that is learnable by the model, after one *gradient descent* iteration with appropriate learning rate, how do the model predictions changes for majority of examples?

(A) Gets close to their actual labels

(B) Gets farther away from their actual labels

(C) Both (a) or (b) may happen

Answer: A

(e) **(4 points)** Consider the following dataset: For each of the following activations, indicate if the model

x_1	0	0	1	1
x_2	0	1	0	1
class	+1	-1	-1	+1

can classify the dataset correctly (Y) or not (N). In each box, write Y or N.

(a) Y ReLU

(b) Y Sigmoid

(c) Y Tanh

(d) N linear

6 Black-box Interpretability (15)

(a) **Shapley Value (9 points)** You want to use Shapley values to explain why a particular image is classified as a 'label', by a neural network. To do so, we first divide the image to 3 super-pixels. The following table shows the probability for the image to be in class 'label' by the neural network, when we input different combinations of super-pixels (Super-pixels kept) to the network and replace the other ones with gray color.

Remember: multinomial coefficient can be calculated as $\binom{n}{m} = n!/m!(n-m)!$*. Note that this is not the exact formula for computation of the Shapley value.*

Super-pixels kept	Probability of 'label'
{}	0.0
{1}	0.4
{2}	0.5
{3}	0.4
{1,2}	0.8
{1,3}	0.5
{2,3}	0.8
{1,2,3}	0.9

(a) **(3 points)** What's the contribution of super-pixel 2 to prediction of 'label' conditioned on (when added to) super-pixels $\{1, 3\}$?

(A) 0.4/3
(B) 0.4/6
(C) 0.3/6
(D) 0.4/3 + 0.4/6 + 0.3/6

Answer: A

(b) **(3 points)** What's the contribution of super-pixel 2 to prediction of 'label' conditioned on (when added to) super-pixel 1?

(A) 0.4/3
(B) 0.4/6
(C) 0.3/6
(D) 0.4/3 + 0.4/6 + 0.3/6

Answer: B

(c) **(3 points)** What's the contribution of super-pixel 2 to prediction of 'label'?

(A) 0.4/3 + 0.5/6
(B) 0.4/6 + 0.5/3
(C) 0.4/6 + 0.4/6 + 0.4/6 + 0.5/6
(D) 0.4/3 + 0.4/3 + 0.4/3 + 0.5/3
(E) 0.4/6 + 0.4/3 + 0.4/3 + 0.5/6
(F) 0.4/3 + 0.4/6 + 0.4/6 + 0.5/3

Answer: F

(b) **LIME (6 points)** Next, we consider using LIME for the above problem. After model selection, we find the following liner model, where X_i models super-pixel i:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2,$$

	β_0	β_1	β_2	β_3	β_4
value	0.1	5	0.1	0.1	5
p-value	0.05	0.1	0.1	0.1	0.01

with the following coefficients:

(4 points) Which of the following statements are correct? You can choose multiple correct answers, but we will deduct 2 points for each incorrect choice.

(A) super-pixel 1 contributes to prediction of 'label' only for this image

(B) super-pixel 1 contributes to prediction of 'label' for all images of this class

(C) super-pixel 2 contributes to prediction of 'label' only for this image

(D) super-pixel 2 contributes to prediction of 'label' for all images of this class

(E) super-pixel 1 or 2 alone do not contributes to prediction of 'label' for this image, but the combination does

(F) super-pixel 1 and 2 together contributes to prediction of 'label' for all images of this class

Answer: E

(c) **(2 points)** When do we use LIME and SHAP?

(A) When the test performance is low and we want to know why

(B) When the test performance is high and we want to make sure model behavior is reasonable

(C) Both (a) and (b) are correct

Answer: C

7 Decision Boundary (10 Points)

Consider the dataset with two classes in the figure below. In each of the plots (a)-(e), one of the following classification methods has been used, and the resulting decision boundary is shown:

(1) Multi-layer Neural Network with linear activation functions

(2) Multi-layer ReLU network

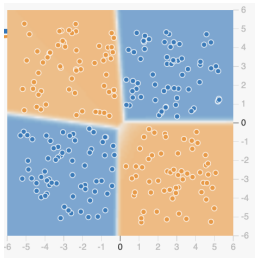
(3) Regularized Multi-layer ReLU network

(4) Multi-layer tanh network

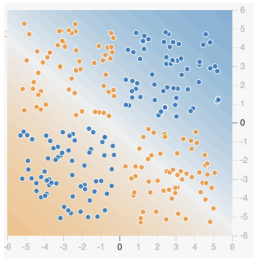
(5) Regularized Multi-layer tanh network

Assign each of the previous methods to exactly one of the following plots (in a one to one correspondence), and explain briefly why you made each assignment. Write your choices (a, b, c ...) in the answer boxes below and your explanation in the corresponding explanation box. There is no partial credit for each part.

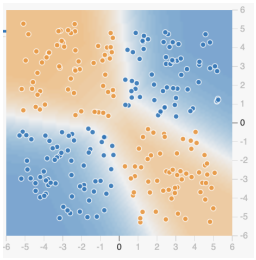
(1) b | Explanation: Decision Boundary is Linear



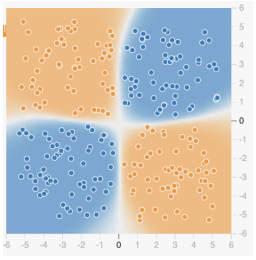
(a)



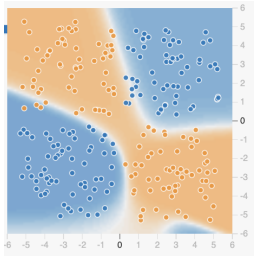
(b)



(c)



(d)



(e)

(2) a	Explanation: Sharp and Piecewise Linear Decision Boundary
(3) c	Explanation: Soft and Piecewise Linear Decision Boundary
(4) d	Explanation: Curved decision boundary.
(5) e	Explanation: Curved Decision Boundary, smoother and less overfit.