Data Science Fundamentals (CS M148) Submit to Gradescope
Released: Feb 25 **Homework 4** Due: March 7, 2pm

1 Interpretation of Coefficients in Linear Regression

Suppose that there is a cholesterol lowering drug that is tested through a clinical trial on a population of men and women. We aim to model the outcome Y (amount of cholesterol lowering) as a linear function of the drug X_1 and sex X_2 using a single linear regression model. We want to model this problem to test the hypothesis that the effect of drug on lowering the cholesterol was different in men compared to women.

(a) **(3 points)** Write the linear regression model.

(b) **(6 points)** Interpret the coefficients of your linear regression model.

(c) **(3 points)** How do you test the hypothesis that the effect of drug on lowering the cholesterol was significantly different in men compared to women?

2 Interpretation of Coefficient in Logistic Regression

We have data from a survey experiment in which respondents were randomly assigned to either the control or the treatment groups. In the survey, respondents were asked to choose whether they prefer A or B. Then, after receiving some stimulus by people in the treatment group, respondents were asked to choose between the same two things: A or B. We want to know if being in the treatment group had an effect on the choice that respondents made in the question. Given that we are working with categorical data, we decide to use a logistic regression model as follows:

Coefficients	Estimate	Std. Error	z value	$Pr(>\|z\|)$
(Intercept)	3.135	1.021	3.070	0.00214
Treatment	-2.309	1.054	-2.190	0.02853
PreferA	-5.150	1.152	-4.472	7.75e-06
Treatment:PreferA	2.850	1.204	2.367	0.01795

(a) **(5 points)** (i) Interpret the Intercept, and (ii) complete the following sentence based on your interpretation: among people who did not prefer A previously and did not receive the treatment, there are _____ persons who prefer A for every such person that prefers B. (iii) Is A popular among such people?

(b) **(5 points)** (i) Interpret the coefficient for Treatment, and (ii) complete the following sentence based on your interpretation: among people who were treated and did not prefer A previously, there are _____ persons who prefer A for every such person who prefers B. (iii) Compare the popularity of A among such people with those considered in part (a).

(c) **(5 points)** (i) Interpret the coefficient for PreferA, and (ii) complete the following sentence based on your interpretation: the baseline odds decreases by a factor _____ when someone preferred A previously. (ii) What does this imply about the treatment?

(d) **(5 points)** (i) Interpret the coefficient for the interaction term, and (ii) and complete the following sentence based on your interpretation: the ratio by which the odds ratio changes between people who preferred A before and that did not is ______.

3 Decision Tree

The following table contains training examples that help predict whether a patient is likely to have a heart attack.

PATIENT ID	CHEST PAIN?	MALE?	SMOKES?	EXERCISES?	HEART ATTACK?
1.	yes	yes	no	yes	yes
2.	yes	yes	yes	no	yes
3.	no	no	yes	no	yes
4.	no	yes	no	yes	no
5.	yes	no	yes	yes	yes
6.	no	yes	yes	yes	no

1. **(15 points)** Use gini gain to construct a minimal decision tree (i.e., stop branching when nodes are pure) that predicts whether or not a patient is likely to have a heart attack. Sow steps of your computation.
2. **(5 points)** Translate your decision tree into a collection of decision rules.

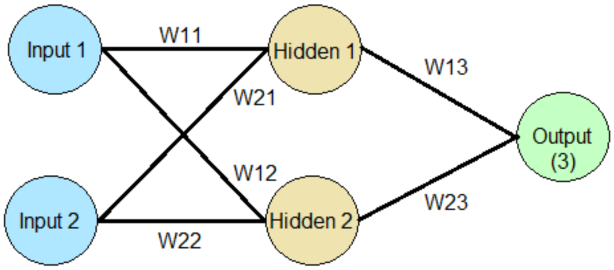
4 Backpropagation

Consider the following network with sigmoid activation functions in the hidden and output neurons, and binary cross entropy loss function. Assume we initialize the weights as follows: $W11 = 1, W12 = 0.5, W21 = 0.1, W22 = 0.2, W13 = 1, W23 = 0.5$. The biases for the hidden nodes are initialized as $b1 = 0.1, b2 = 0.1$, and for the the output node is initialized as $b3 = 0.5$.

(a) **(4 points)** Forward pass: For input 1 = 0.1, input 2 = 0.2, and label = 1, calculate the value of the output and the loss.

(b) **(15 points)** Backward pass: Calculate the derivative of the loss w.r.t $W11$. What's the value of the derivative for the initialized weights, input 1 = 0.1, input 2 = 0.2, and label = 1?

(c) **(6 points)** How do you update $W11$ using gradient descent, based on the derivative you derived in the previous part and learning rate $\eta = 0.1$? What is the value of the loss using the updated $W11$? How did it change after the update?



5 Clustering

Consider the dataset in Figure 1 with two clusters C_1 and C_2. Each cluster lies in a ball with radius 1 centered at $c_1 = (-1 - d, 0)$ and $c_2 = (1 + d, 0)$. Suppose we want to use K-means to do the clustering. Recall that K-means algorithm will perform the following steps until convergence: (1) initialize cluster centers $u^{(i)}$ and the number of clusters K; (2) compute distances $\|x^{(i)} - u^{(i)}\|_2$; (3) assign points to nearest cluster center; (4) update cluster centers $u^{(j)} = \frac{1}{N_j} \sum_{x_i \in C_j} x_i$.

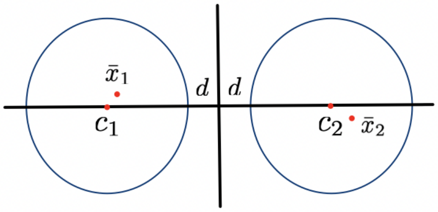
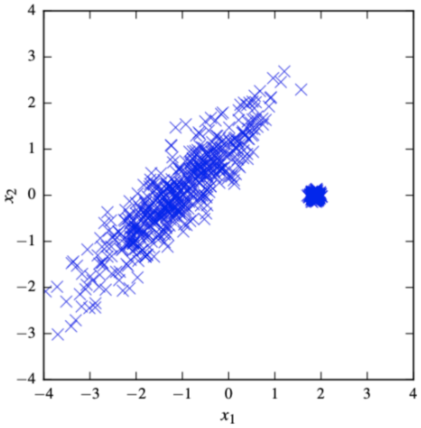


Figure 1: Dataset for K-means.

(a) **(5 points)** Suppose we have the mean of the cluster C_k as $\bar{x}_k = \frac{1}{N_k} \sum_{x_i \in C_k} x_i$, where N_k is the number of points in C_k, and $\|\bar{x}_k - c_k\|_2 \leq d$ for $k = 1, 2$, as shown in Figure 1. If we initialize our K-means with $K = 2$ and $u^{(k)} = c_k$, show that our method will converge after the first iteration. (*The convergence means the center and clustering will not change.*)

(b) **(5 points)** If d is large enough, give a simple explanation that any initialization $u^{(k)} \in C_k$ will converge after the first iteration.

(c) **(5 points)** K-means is sensitive to outliers that pull the cluster centers towards them. Which clustering algorithm can you use if you expect to have outliers in your data? In the figure below, draw (approximately) the clusters you get from K-means clustering with K=2, and compare it with the clusters your get from your choice of clustering algorithm. Use appropriate parameters for your choice of clustering algorithm in a way that you get 2 clusters, and draw the clusters in each case.



6 Decision Boundary

Consider the classification problems with two classes, which are illustrated by circles and crosses in the plots below. In each of the plots, one of the following classification methods has been used, and the resulting decision boundary is shown:

(1) **(1 points)** Linear SVM

(2) **(1 points)** Kernelized SVM (Polynomial kernel of order 2)

(3) **(1 points)** Perceptron

(4) **(1 points)** Logistic Regression

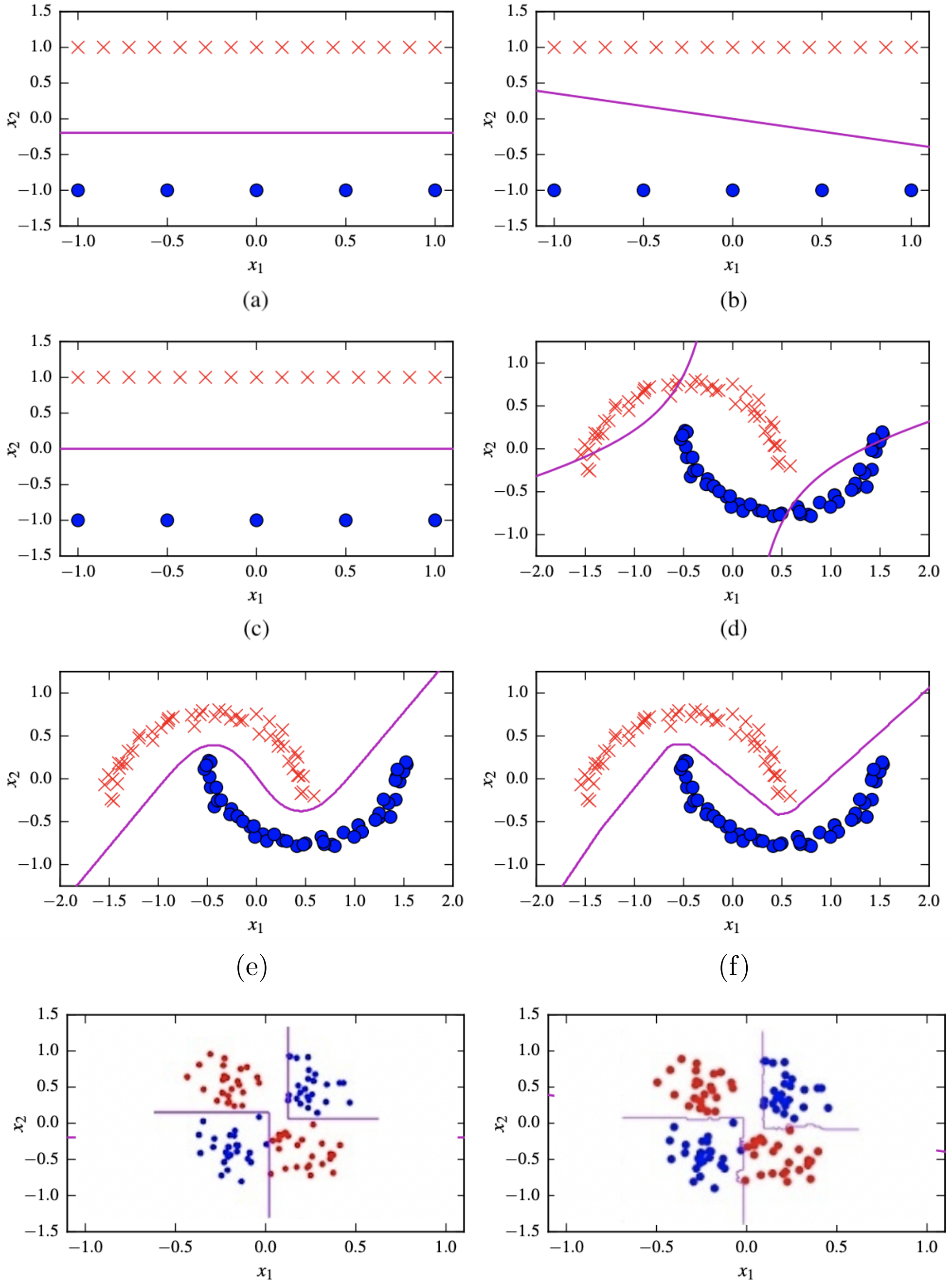
(5) **(1 points)** Decision Tree

(6) **(1 points)** Random Forest

(7) **(1 points)** Neural Network (1 hidden layer with 10 ReLU)

(8) **(1 points)** Neural Network (1 hidden layer with 10 tanh units)

Assign each of the previous methods to exactly one of the following plots (in a one to one correspondence) by annotating the plots with the respective letters, and **explain briefly** why did you make each assignment.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)