Michigan State University **Sep 27, 2023**
CMSE381 - Data Science

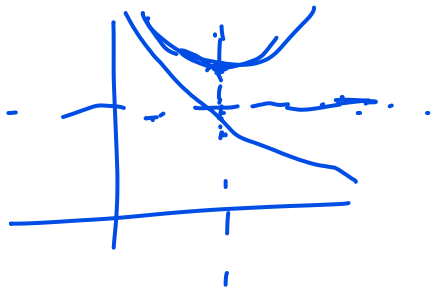
CMSE381 - Midterm #1

I will adhere to the Spartan Code of Honor in completing this assignment.

Signed: _____

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (3:00-4:20pm) to complete the exam.
3. This exam is closed book. Unless otherwise specified, you may use any calculator as long as there is no internet connection.
4. You may use one cheat sheet to the test. This is one 8.5"x11" sheet of paper. It must be handwritten and must be your own work. Photocopies and computer print outs are not allowed. You will turn in your sheet with your test, so make sure your name is on it.
5. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. $\boxed{\text{BOX}}$ your answers. Partial credit will be given where warranted.
6. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P

1. (16 points)



(a) Logistic regression is used for regression.

TRUE FALSE

(b) The best model will _____ have training error below the irreduceable error.

Always Sometimes Never

(c) The best model will _____ have test error below the irreduceable error.

Always Sometimes Never

(d) Increasing your model flexibility always results in a better model.

TRUE FALSE

(e) A logistic regression model is set up so that the logg odds are linear.

TRUE FALSE

(f) Circle all of the following that would represent a qualitative variable.

Age Year Dog_breed Country_of_origin

Student_(True/False) Weight Speed MPG

(g) What equation would you use to evaluate the result of a regression model?

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

(h) What equation would you use to evaluate the result of a classification model?

$$\text{err rate} = \frac{1}{n}\sum_{i=1}^{n} 1\{y_i \neq \hat{y}_i\}$$

2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.

   (a) List all input variables.

   (b) List all output variables.

   amount of dog food

   (c) Is this a regression or classification problem? How do you know?

   (d) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. How would you encode the `breed` data for use in the model?

	X_{i1}	X_{i2}
H	1	0
T	0	1
S	0	0

3. (15 points)

(a) Explain in 1-2 sentences the meaning of the "bias-variance tradeoff".

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.

4. (12 points)

(a) What is the Bayes classifier?

Classifies that pick the most likely class based on the $P(Y=j \mid X=x_0)$

(b) What is the Bayes decision boundary?

location where the Bayes classifier switch the prediction

(c) The table below provides a training data set containing seven observations, three predictors, and one qualitative response variable. I have also included the distance from each observation to the test point $X_1 = X_2 = X_3 = 0$. If we use k-nearest neighbors classification with $k = 3$, what is the prediction for $X_1 = X_2 = X_3 = 0$?

Obs.	X_1	X_2	X_3	Y	Distance
1	-2	1	-1	Chicken	2.45
2	0	1	0	Duck	1.00
3	1	0	2	Chicken	2.24
4	-1	2	3	Duck	3.74
5	1	3	-1	Chicken	3.32
6	1	1	1	Chicken	1.73
7	-1	2	2	Chicken	3

5. (12 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$X_1 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$
$$X_2 = \texttt{Number\_of\_sentences}$$

and am training a logistic model to predict

$$Pr(Y = \texttt{spam} \mid X_1, X_2).$$

(a) Write down the equation for the model you would train, using our standard notation with β_i's.

$$P(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}{1 + \ldots}$$

(b) If my trained model used $\beta_0 = -8.1$, $\beta_1 = 7.8$, and $\beta_2 = 0.3$, what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .

$X_2 = 5$ $X_1 = 1$

(c) How would you change the encoding of your model if you were trying to predict whether your incoming email was from the set {`Spam`, `Not_important`, `Urgent`}.

Baseline softmax encoding

6. (20 Points) In our `diabetes` data set, we are predicting `target`, a quantitative measure of disease progression one year after baseline. We are training a linear model to predict `target` from `age`, `bp`, and `s1`.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	152.1335	3.276	46.433	0.000	145.694	158.573
age	37.6853	74.559	0.505	0.614	-108.852	184.223
bp	660.0505	74.208	8.895	0.000	514.203	805.898
s1	173.4155	72.400	2.395	0.017	31.122	315.709

Dep. Variable:	target	**R-squared:**	0.207
Model:	OLS	**Adj. R-squared:**	0.202
Method:	Least Squares	**F-statistic:**	38.13
Date:	Sun, 24 Sep 2023	**Prob (F-statistic):**	6.56e-22
Time:	22:16:55	**Log-Likelihood:**	-2495.9
No. Observations:	442	**AIC:**	5000.
Df Residuals:	438	**BIC:**	5016.
Df Model:	3		
Covariance Type:	nonrobust		

< 0.05

(a) What is the equation of the learned model?

$y =$

(b) Which variable are we least confident in and why?

Continued on next page...

...Continued from previous page

(c) What are the null and alternative hypotheses for the hypothesis test we would use the F-statistic for?

H_0. $\beta_1 = \beta_2 = \beta_3 = 0$

H_A. at least one β nonzero

(d) What would be the conclusion of that hypothesis test? Why?

reject H_0

as $p < 0.05$

(e) Give an approximate 95% confidence interval for ~~radio~~. bp

7. (10 Points) In our familiar `auto` data set, a student is predicting `mpg` from `origin` using a linear model. Recall that `origin` labels the country of origin of the car, and takes values 1 for American, 2 for European, or 3 for Japanese.

   Here's the head of the data set.

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino

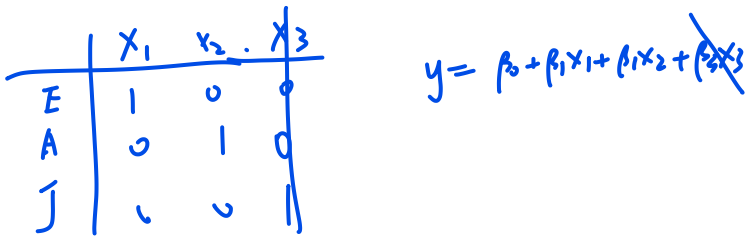
   The student runs the following code to create their model and gets the following output.

```
1  import statsmodels.formula.api as smf
2  import pandas as pd
3  df= pd.read_csv('Auto.csv')
4
5  est = smf.ols(formula='mpg ~ origin', data=df).fit()
6  est.summary().tables[1]
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	14.8623	0.716	20.760	0.000	13.455	16.270
origin	5.4967	0.405	13.564	0.000	4.700	6.293

   (a) What is wrong with this model?

   (b) How would you fix it?



Scrap Paper