CS-482 Machine Learning
Linear Models and Metrics
Chapter 3
Homework Questions

Linear Regression

1. A regression analysis relating test scores (Y) to training hours (X) produced the following fitted question: $\hat{y} = 25 - 0.5x$.

(a) What is the fitted value of the response variable corresponding to $x = 7$?
(b) What is the residual corresponding to the data point with $x = 3$ and $y = 30$?
(c) If x increases 3 units, how does $\hat{y}$ change?
(d) An additional test score is to obtained for a new observation at $x = 6$. Would the test score for the new observation necessarily be 22? Explain.
(e) The error sums of squares (SSE) for this model was found to be 7. If there were $n = 16$ observations, provide the best estimate for σ^2.

2. Explain the different between the following two equations which are linear models for predicting Y given X.
$$\hat{Y} = b_0 + b_1 X$$
$$Y = \beta_0 + \beta_1 X + \epsilon.$$

3. Given the following predictor matrix P with 2 predictors and sample size of 3,
$$P = \begin{bmatrix} 2 & 3 \\ 4 & 2 \\ 9 & 1 \end{bmatrix}$$

and the target y = $\begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}$

a) Find the parameters of the linear model β by solving the equation

b) Predict the value of $\hat{y}$ for $x_i = [\,4\ 2\,]$ $x_j = [\,1\ 5\,]$
c) Find R^2, the correlation between predicted values and target values the model developed in a) $\beta = (X^T . X)^{-1} X^T y$
d) Find the correlation coefficient r between the two predictors.

4. State one reason for using Ridge regression and one reason for using lasso regression.

5. Given the following table

Predicted	Observed	
	Cats	Dogs
Cats	8	7
Dogs	5	3

Compute Accuracy, Recall, Precision and Specificity and f-statistic

6. A classification model with four classes predicted continuous value prediction of (.6 .4 .3 .2) for the probability of the four classes. Convert these using softmax to probability values that add to 1.

7. Given the following predictions of a binary classification, draw the calibration plot using 4 equal bins.

	Observed	Predicted (A,B)
1	A	(.4, .6)
2	B	(.5, .5)
3	A	(.2 .8)

4 A   (.8 .2)
5 B   (.3 .7)
6 A   (.1 .9)
7 A   (.2 .8)
8 B   (.4 .6)
9 A   (.6 .4)
10 B   (.7 .3)

8. Given the following classification data
a) find the log odds values for each sample.
b) Find the $Pr\,[Y = C_l]$
1.Number of times pregnant.   2. Plasma glucose concentration in a 2 hours in an oral glucose tolerance test. 3. Diastolic blood pressure (mm Hg). 4. Triceps skinfold thickness (mm). 5. 2-Hour serum insulin (mu U/ml). 6. Body mass index (weight in kg/(height in m)^2). 7. Diabetes pedigree function. 8. Age (years). 9. Class variable (0 or 1).

1	2	3	4	5	6	7	8	9
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1