INFO 7390 – ADS
Practice Exam One Solutions

Student Name: _____
Professor: Nik Bear Brown

Rules:
1. NO COMPUTER, NO PHONE, NO DISCUSSION or SHARING.
2. Ask if you don't understand a question.
3. You may use one 8½"×11" sheets of notes (you may use both sides, written or printed as small as you like).
4. Time allowed. Until end of class.
5. Bring pen/pencil. The exam will be written on paper.

Q1 (5 Points)

What is a low bias model? What are its advantages/disadvantages? How do we counter act its disadvantages?

Solution:

A low bias model is a flexible model like a decision tree or spline. Unlike linear model a high bias model.

Low bias overfits.

It can fit complex data.

We counter act its disadvantages, by testing it versus out of sample data.

Q2 (5 Points) Assume regression is being used to predict whether a student will get drunk or not. The dependent variable is **drunk,** which indicates drunk or not. Assume the only independent variable is **beers,** which indicates the number of beers consumed. The stats for the fit are shown in the table below.

drunk	Coef.	Std. Err.
beers	0.7	0.035
intercept	-2.5	0.1

A. Write an equation that describes the model.
B. Is the coefficient *beers* significant? How does one interpret the meaning of its value?
C. Is the coefficient *intercept* significant? How does one interpret the meaning of its value?
D. What is the probability of getting drunk after 2 beers?
E. What is the likelihood of getting drunk after no beers?

Solution:

A.

$\text{logit}(p) = \beta_0 + \beta_1 * \textbf{beers} + \textbf{e}$

Any form of this equation is fine. (Like the one a little further down in this answer in terms or p rather than logit(p))

B. The coefficient beers is significant as the z-score is 20 SD above the mean.

0.7/0.035 = 20

The coefficient *intercept* is significant as the z-score is 25 SD below the mean

2.5/0.1 = -25

How do we interpret the coefficient for beers? The coefficient and intercept estimates give us the following equation:

$\text{log}(p/(1-p)) = \text{logit}(p) = -2.5 + 0.7 * \textbf{beers}$

The coefficient for **beers** is the difference in the log odds. In other words, for a one-unit increase in the beers score, the expected change in log odds is .7

The likelihood of getting drunk after no beers is just the *intercept*

So just exponentiate the log- likelihood of -2.5 or exp(−2.5) which is very low or 0.08

Logistic Regression

Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using `balance` to predict `default`. Logistic regression uses the form

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

($e \approx 2.71828$ is a mathematical constant [Euler's number.])
It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.

P can be computed from the regression equation for a given value of X.

$$P = \frac{\exp(a+bX)}{1+\exp(a+bx)} = \frac{e^{a+bx}}{1+e^{a+bx}}$$

Q3 (5 Points) Write pseudocode for the bootstrapping algorithm.

Solution:

Choose any number k bootstrap samples.

With k bootstrap samples of same size has original data with replacement.

Or

```
for i in bootstraps:
      sample = select_sample_with_replacement(data)
```

Q4 (5 Points) Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

Solution:

Parametric models have parameters that are fit from the data.

We can interpret the parameters.

We are often limited by the parameteric structure of the model.

Q5 (5 Points) What is the difference between Ridge and Lasso regression? Why use Ridge or Lasso regression? Why would I use one or the other? Are there hyperparameters in use Ridge or Lasso regression, and if so, how does one determine their value?

Solution:

Ridge regression is squaring the coeffecients and Lasso takes absolute value.

$$\hat{\beta}^{\text{ridge}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

Why use Ridge or Lasso regression?

For regularization

Why would I use one or the other?

Ridge and lasso regression allow you to regularize ("shrink") coefficients. This means that the estimated coefficients are pushed towards 0, to make them work better on out of sample data.

Lasso pushed coefficients to 0, and Ridge near 0.

Are there hyperparameters in use Ridge or Lasso regression, and if so, how does one determine their value?

Cross-validate.

Q6 (5 Points) How does one adjust the support in a Support Vector Machine? How does one adjust the bias in a Support Vector Machine other than changing the kernel? Are their hyperparameters that adjust the support and bias? If so, how does one determine their values?

Solution:

How does one adjust the support in a Support Vector Machine?
Adjust the support in a Support Vector Machine by adjusting the budget C.

How does one adjust the bias in a Support Vector Machine other than changing the kernel?
Adjusting the bias by adjusting gamma.

Are their hyperparameters that adjust the support and bias?
Yes. The budget and gamma.

If so, how does one determine their values?
Cross-validate.

Q7 (5 Points) Write the Ridge regression equation. Are there coefficients not effected by the tuning parameter?

Solution:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$
$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

Yes the intercept is not effected.

Q8 (5 Points) What is the equation for multiple logistic regression with two independent variables?

Solution:

Multiple logistic regression when you have one nominal variable and two or more measurement variables. This is not the same as multi-class logistic regression. But based on the responses I feel that I didn't make this clear enough in class.

Because I feel the distinction was unclear in the lecture either the equation for multiple (multivariate)

Equation for multiple logistic regression

$\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \ldots + \beta_n XN + e$

Equation for multiple logistic regression with two independent variables

$\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 X1 + \beta_2 X2 + e$

Equation for multi-class logistic regression (NOT multiple logistic regression)

Logistic regression with more than two classes

So far we have discussed logistic regression with two classes. It is easily generalized to more than two classes. One version (used in the R package `glmnet`) has the symmetric form

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+...+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+...+\beta_{p\ell}X_p}}$$

Here there is a linear function for *each* class.

Q9 (5 Points) Describe an interaction variable in regression. How does one create them? How does one know whether their effect is significant?

Solution:

It is a product term. Multiply terms. Check the z/t-score or pvalue

Q10 (5 Points) Create an algorithm for aggregation of base models in bagging that uses another machine learning model rather than numerical aggregation.

Solution:

Take the output of the models as independent variables and use a surrogate model to predict the known output.

Say an algorithm outputs algorithm one prediction, algorithm two prediction, and algorithm three prediction and the actual output is target y. Then any supervised learning algorithm could be used to take the three predictions as independent variables to predict the dependent variable.

Q11 (5 Points) Write pseudocode for the K-Means Clustering algorithm.

Solution:

Decide a k.
Place place centroids.
Assign all points to nearest centroid
Move centroid to center of cluster
Repeat until no points reassigned

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {

   For every i, set
   $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each j, set
   $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

   }

Q12 (5 Points) Assume one wants to use the three-class categorical variable high/medium/low as an independent variable in linear regression. How can we encode it?

Write an equation that describes the model. Assume we are fitting the intercept and one other continuous independent variable, and our dependent variable is called 'y'.

Solution:

One would pick a base class and create dummy variables for the other two.

The multiple linear regression equation is as follows:

$y_i=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\epsilon_i$.

Where two of the slopes are dummy variables and the third is continuous independent variable and β_0 is the intercept.

Credit will be given for one-hot encoding but the multiple linear regression equation must be as follows:

$y_i=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\epsilon_i$.

Where three of the slopes are one-hot variables and the fourth is continuous independent variable and β0 is the intercept.

Q13 (5 Points) What is a standardized regression coefficient? How do we calculate them?

Solution:

They are coefficients on the same scale.
Subtract mean and divide by the standard deviation.

Q14 (5 Points) Describe k-fold cross validation in pseudocode.

Solution:

Divide k-times. Hold one out train on other k-1 folds. Do this k-times and average the results.

Q15 (5 Points) Describe the difference between bagging, boosting and stacking? Which are ensemble methods?

Solution:

bagging
Create k bootstrap samples
Train models on each bootstrap sample
Aggregate the results

Boosting
Start like bagging but hold out misses and include the further sample

Stacking
Ensembling different algorithms

Q16 (5 Points) Is k-means a parametric or a non-parametric statistical learning approach. Why or why not?

Solution:

non-parametric it has no parameters just the hyperparameter k.

Q17 (5 Points) Write the equation for multiple linear regression with three continuous independent variables and one categorical independent variable with two classes. Are there any assumptions of the error model?

Solution:

We have three continuous independent variables and one dummy independent variable (two classes minus one).

One would pick a base class and create a dummy variable and the number of classes for the one categorical independent variable.

If one picks a two-class categorical independent variable, then there will be 2-1 or 1 dummy variable.

The multiple linear regression equation is as follows (for a 2-class categorical independent variable):

$$yi=\beta0+\beta1x1+\beta2x2+\beta3x3+\beta4x4+\epsilon i.$$

Where one of the slopes is the dummy variable and the three are the continuous independent variable and β0 is the intercept.

Error is assumed to be normal and homodescatic.

Q18 (5 Points) What is the difference between a t-test and Z-test? How are z-scores calculated?

Solution:
t-test uses a t-distribution
Z-test uses a normal distribution

Subtract mean/divide by standard error

The formula that we will use is as follows: $z = (x - \mu)/ \sigma$

The description of each part of the formula is:

- x is the value of our variable
- μ is the value of our population mean.
- σ is the value of the population standard deviation.
- z is the z-score.

Q19 (5 Points) Assume you built a linear regression model showing 95% confidence interval for a coefficient. Does it mean that there is a 95% chance that the model coefficient is a real (i.e. significant) effect?

Solution:

No we need a hypothesis test (i.e. z-score/p-value)

Q20 (5 Points) You are asked to build a classification model about meteorites impact with Earth (important project for human civilization). After preliminary analysis, you get 97% accuracy. Is this a good result? Why or why not?

Solution:

No very rare event. We can get a 99.99999+% accuracy by just predicting no meteorite impact.