INFO 6105 – Data Sci Eng Methods
Exam Two 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)

When should you use classification over regression? What metrics are appropriate classification and regression? What is the null model in regression? What is a confusion matrix? Draw an example confusion matrix.

Solution:

When our dependent variable is categorical.

Metrics classification
- AUC, logloss, mean per class error, accuracy. etc

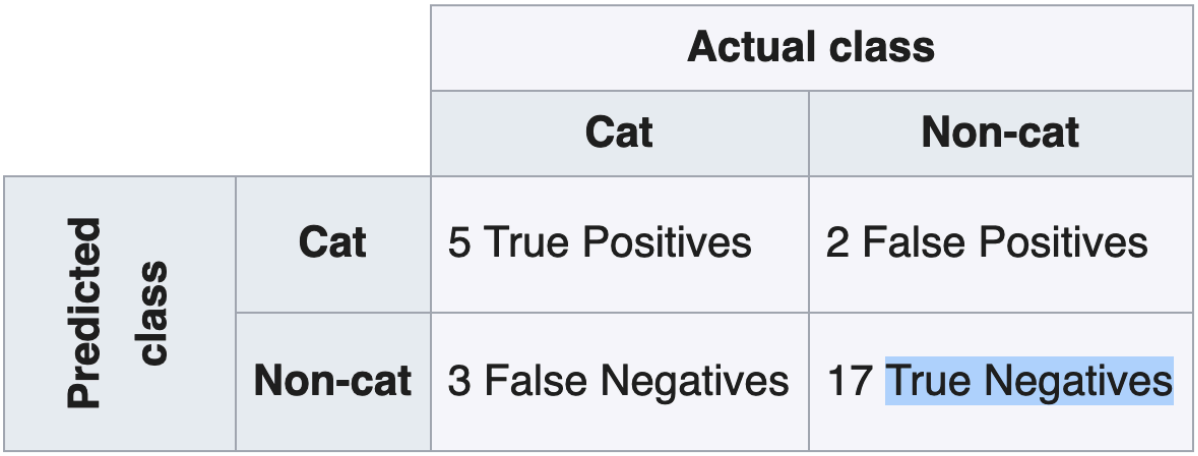
Metrics regression

Measures of distance
Residual deviance, logloss,RSME. MAE, etc.

What is the null model in regression?
The mean

What is a confusion matrix?

A confusion matrix, also known as an error matrix, is a table of True positives, False positives, False Negatives and True Negatives.



		Actual class	
		Cat	Non-cat
Predicted class	Cat	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

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

```
  pass       | Coef.     | Std. Err.  |
------------+-----------+------------+
  hours      | 0.2        | 0.01        |
  intercept | -2.5       | 0.1         |
```

A. Write an equation that describes the model.
B. Is the coefficient *hours* 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 passing after 2 study hours?
E. What is the likelihood of passing with no study hours.

Solution:

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

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 hours is significant as the z-score is 20 SD above the mean.

0.2/0.01 = 20

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

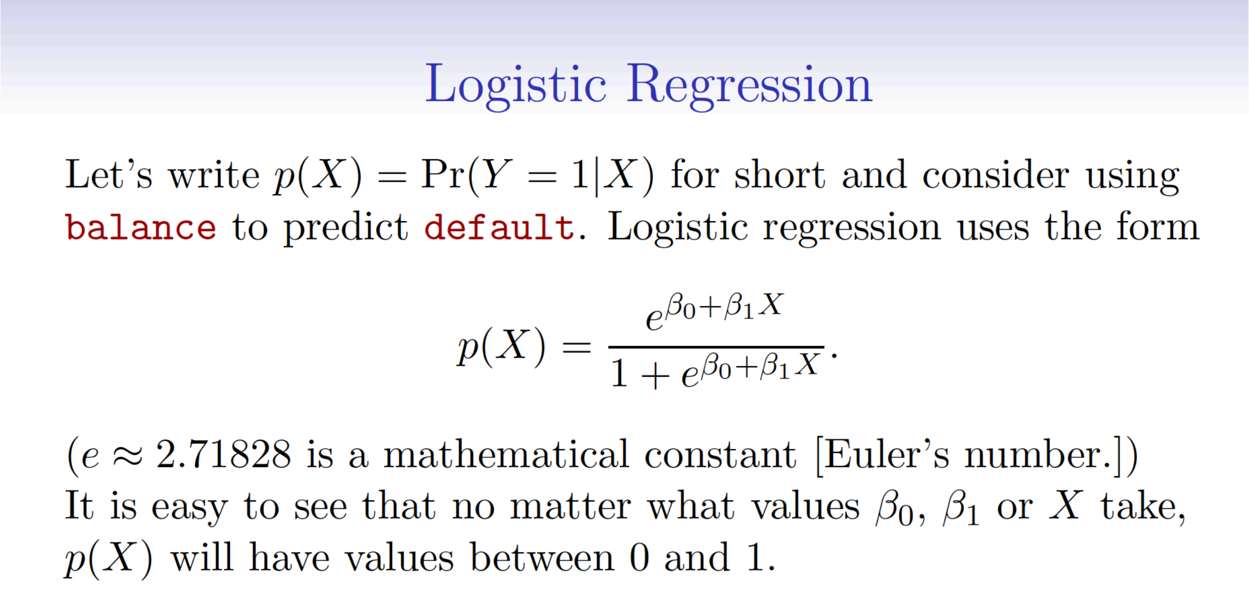
-2.5/0.1 = -25

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

$\log(p/(1-p)) = \text{logit}(p) = -2.5 + 0.2 * \textbf{hours} = -2.5 + 0.2 * \textbf{2}$

The coefficient for **hours** is the difference in the log odds. In other words, for a one-unit increase in the hours score, the expected change in log odds is .4 (2*.2)

The likelihood of passing after no hours 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 boosting algorithm.

Solution:

Starts with bootstrap sample

Sample of same size as original data with replacement.

Train model. Keep misses in second sample, rest fill from original data with replacement.

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 the 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.

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

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:

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

Yes the intercept is not effected.

Q8 (5 Points) Define model bias. Define model variance. Explain the Bias-Variance Tradeoff.

.

Solution:

model bias is error on training data
model variance is error on testing data
Bias-Variance Tradeoff – after a point when we reduce bias we will increase variance.

Q9 (5 Points) Define a hyperparameter. How does one determine their values? Give an example of two hyperparameters used in Random Forests.

Solution:

Hyperparameter is a tuning parameter specified BEFORE a model is fit.

Determine their values by trying some and cross-validating.

hyperparameters used in Random Forests – number of trees, max depth

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) What is the ROC Curve and what is AUC (a.k.a. AUROC)? When would one use one or the other? Is AUC more useful for classification or regression?

Solution:

The ROC curve is a plot with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

AUC is a number the represents the area under the ROC curve.

We compare models with a single number we can see more detail with a ROC curve.

AUC is used for binary classification

Q12 (5 Points) Why do ensembles typically have higher scores than individual models?

Solution:

Because they adjust for biases in individual models.

Q13 (5 Points) Assume you created a recommender algorithm that has a success rate of 2.3% over another algorithm that has a success rate of 2.1%. How can you prove that the improvement you've brought to the algorithm is really an improvement over not doing anything?

Solution:

A hypothesis test.

.

Q14 (5 Points) How can you ensure that you are not overfitting with a particular model? Explain the role of regularization in overfitting.

Solution:

We test on out of sample data to prevent overfitting.

Regularization is used to reduce the model complexity which helps prevent overfitting.

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) Assume logistic regression and SVM models have the same accuracy and computational performance on some data. Name at least two advantages logistic regression would have over SVM in this case.

Solution:

logistic regression allows us to generate probabilities

logistic regression coeffecients are interpretable

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

Solution:

We

We have two continuous independent variables and two dummy independent variables (three 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):

$yi=\beta 0+\beta 1x1+\beta 2x2+\beta 3x3+\beta 4x4+\epsilon i.$

Error is assumed to be normal and homodescatic.

Q18 (5 Points) What are the parameters of a Normal Distribution? Can one create a Normal Distribution from uniform distributions?

Solution:

parameters of a Normal Distribution

mean and variance (or SD)

Can one create a Normal Distribution from uniform distributions?

Yes, by summing iid samples (ie. Central limit theorem)

Q19 (5 Points) Why might it be preferable to include fewer features (predictors) over many? Would we expect the evaluation metric to improve with fewer features (predictors)?

Solution:

fewer features is more interpretable and less likely to overfit

Would we expect the evaluation metric to improve with fewer features (predictors)?
No. At best it would be the same.

Q20 (5 Points) Bootstrapping is a statistical method for estimating properties of an estimator (such as variance, confidence intervals) by measuring those properties when sampling from an approximating distribution.

Describe the algorithm in detail. Name another algorithm that uses bootstrapping.

Solution:

Bootstrapping

If your data has k rows, each bootstrap sample generated randomly samples the original *with replacement* the same size as the original dataset (k rows). We can generate as many bootstrap samples as we wish.

Any algorithm that uses bagging, such as a Random Forest, use bootstrapping. In fact, the name “bagging” stands for “bootstrap aggregation”. Or just mentioning bagging or boosting is fine.

-2 if you don’t mention *with replacement.*